PB81-188930

Evaluation of Statistical Procedures for Grading Fuel Efficient Engine Oils

Falcon Research and Development Co.,
Buffalo, NY

Prepared for

Environmental Protection Agency
Ann Arbor, MI

Nov 80

U.S. Department of Commerce
National Technical Information Service
NTIS

EPA 460/3-81-012

FALCON RESEARCH

Falcon Research & Development Co.
A Subsidiary of Whittaker Corporation
One American Drive
Buffalo, New York 14225
716/632-4932

Whittaker

Evaluation of Statistical Procedures for Grading Fuel Efficient Engine Oils

Report 3520-2/BUF-40

Prepared Under

Contract 68-03-2835
Task Order #2

Prepared for

Environmental Protection Agency
Ann Arbor, MI 48105

Prepared by: S. Kaufman
H. T. McAdams
N. Morse

Date: November 1980

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPI PB81-188930
EPA 460/3-81-012		
4. TITLE AND SUBTITLE Evaluation of Statistical Procedures for Grading Fuel Efficient Engine Oils		5. REPORT DATE November 1980 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) S. Kaufman, H.T. McAdams, and N. Morse		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Falcon research and Development Co. One American Drive Buffalo, New York 14225		10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO. 68-03-2835
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency Ann Arbor, Michigan 48105		13. TYPE OF REPORT AND PERIOD COVERED Final 14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

Reproduced from best available copy.

16. ABSTRACT

In March 1980 the EPA distributed for industry comment a draft "EPA recommended Practice for Evaluating, Grading, and Labeling the Fuel Efficiency of Motor Vehicle Engine Oils." EPA requested Falcon Research and Development Co. to "Analyze comments received...with regard to the recommended statistical procedure."

As a follow-up effort, Falcon was requested to provide an independent assessment of certain aspects of the proposed statistical procedure. Particular areas identified for consideration were: 1) comparative accuracy of carryover and noncarryover effect procedures, 2) means for encouraging (rewarding) accurate testing, 3) dealing with outliers in fuel economy test data and 4) the impact of variability in oil effects across car models on the meaningful grading of oils. An investigation of these areas, and cursory examination of some additional topics form the subject of this report.

PRODUCT OF:
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Fuel consumption Fuel oil Motor Vehicles Marking Evaluation Statistical analysis	Fuel economy	

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
RELEASE UNLIMITED	UNCLASSIFIED	78
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (9-73)

TABLE OF CONTENTS

Section	Title	Page
1.0	INTRODUCTION	1
2.0	SUMMARY AND CONCLUSIONS	2
3.0	FUEL EFFICIENCY RATING ERROR ANALYSIS	5
	3.1 Non-Carryover Procedure	5
	3.2 Carryover Procedure	8
	3.3 Comparative Evaluation	9
4.0	MODEL-TO-MODEL VARIATION IN FUEL EFFICIENCY OF OILS	12
	4.1 Grading Policies Under Model Differences	12
	4.2 Statistical Test for Model Differences	14
	4.3 Rationale for Selection of Test Models	14
5.0	REQUIRED NUMBER OF TEST VEHICLES	16
	5.1 Grading Accuracy Requirements	16
	5.2 Statistical Test of Model Homogeneity	19
6.0	STATISTICAL STRATEGIES IN FUEL ECONOMY MEASUREMENT	22
	6.1 Non-Carryover Procedure	22
	6.1.1 Error Reduction Through Replication	22
	6.1.2 Robust Estimation	23
	6.2 Carryover Procedure	25
	6.2.1 Dominance of Extended Mileage Accumulation Variability	25
	6.2.2 Proposed Alternative Experimental Design	26
	6.3 Concluding Remarks	27
7.0	SCALING ASSUMPTIONS IMPLICIT IN SPECIFIC FUEL EFFICIENCY RATING MEASURES	29

TABLE OF CONTENTS (Continued)

Section	Title	Page
Appendix A	ENGINE OIL FUEL EFFICIENCY GRADING UNDER MEASUREMENT ERROR AND BETWEEN-MODEL DIFFERENCES	A-1
Appendix B	ACCOMMODATING SPURIOUS OBSERVATIONS IN FUEL ECONOMY MEASUREMENTS	B-1
Appendix C	"BEST TWO OUT OF THREE" PROCEDURE	C-1
Appendix D	BEHAVIOR OF EPA REPEATIBILITY TEST	D-1

1.0 INTRODUCTION

In March 1980, the Environmental Protection Agency (EPA) distributed for industry comment a draft "EPA Recommended Practice for Evaluating, Grading, and Labeling the Fuel Efficiency of Motor Vehicle Engine Oils." Following receipt of responses, EPA requested Falcon Research and Development Company to "Analyze comments received...with regard to the recommended statistical procedure." This analysis was completed and documented in June 1980.*

As a follow-on effort, Falcon was requested to provide an independent assessment of certain aspects of the proposed statistical procedure, extending beyond evaluation of just those issues which evolved from the industry review. Particular areas identified for consideration were: (1) comparative accuracy of carryover and noncarryover effect procedures, (2) means for encouraging (rewarding) accurate testing, (3) dealing with outliers in fuel economy test data, and (4) the impact of variability in oil effects across car models on the meaningful grading of oils. An investigation of these areas, and cursory examination of some additional topics, form the subject of this report.

---

* H. T. McAdams, S. Kaufman, and N. Morse, "Analysis of Industry Comments on EPA Recommended Practice for Fuel Efficient Oils," Falcon Research and Development Company, Report 3520-2/BUF-36, June 13, 1980. For convenience this report will henceforth be referred to as Falcon's "Analysis of Industry Comments."

2.0 SUMMARY AND CONCLUSIONS

Section 3 recapitulates prior ASTM and Falcon error analyses of fuel efficiency rating estimation by non-carryover (NCO) and carryover (CO) procedures. The possibility of a carryover effect bias error occurring in the NCO procedure is explicitly considered. Experimental evidence regarding the effectiveness of special flush procedures to minimize such bias in carryover effect oils is not currently available, so that relative evaluation of NCO and CO procedures is not yet possible. However, the potential feasibility of the former, assuming demonstration of a special flush capability, is reaffirmed, whereas the CO procedure, in light of the best available component error estimates, appears doomed to require an exorbitantly large number of vehicle replications to achieve acceptable probabilities of correct grading.

Section 4 introduces the additional problem of potential model-to-model variability in fuel-efficient oil effects. It is observed that if such variations are significant in an oil, i.e., comparable to the grade spacings or greater, then there seems to be no completely satisfactory rule for determining a grade designation for that oil. In any event, in the absence of sufficient experimental evidence that such variabilities are insignificant, it is necessary to determine by statistical test for each candidate oil whether it could be characterized by a single fuel efficiency rating over all models or alternatively whether it is necessary to separately estimate individual model-specific ratings.

The rationale for the five particular models selected for testing is questioned. The point is then made that if it were possible to stratify all vehicle models into a relatively small number of homogeneous classes with respect to oil fuel efficiency effect (on the basis of fundamental physico-chemical properties), then test model selection could be made with less arbitrariness.

In Section 5 the full implications of the issues discussed in the previous two sections are carried through with respect to requirements on the numbers of vehicles for testing a given candidate oil. It is shown that the statistical test for homogeneity across models imposes the severest requirements. On bracketing with two reasonable alternative levels of performance of this test, the determination is made that for the NCO procedure from 65 to 200 vehicles are required, whereas for the CO procedure the comparable requirements range from 1660 and 4650. Even if the problem of model-to-model variability were completely resolved, in order to achieve a probability of misgrading no higher than 10% requires 60 vehicles under the NCO procedure and over 1000 vehicles under the CO procedure. If the 10% standard is relaxed to 20%, the requirements reduce to 20 and 300 vehicles, respectively. It must be strongly emphasized that these numerical results rest on the error component estimates stated in the ASTM response. In particular, the dominant error term which very markedly penalizes the CO procedure is the 2.34% estimate for the standard deviation in car mpg variations about a linear trend with extended mileage accumulation. Downward revision of these errors could significantly alter our conclusion. With this caveat in mind it is concluded that the CO procedure is impractical and that even the NCO procedure imposes a substantial test load for adequately sharp results. The basic problem is seen to be the smallness of the effects to be discerned relative to uncertainties inherent in vehicle-oil performance and the measurement process.

Section 6 is primarily concerned with statistical strategies of measurement replication for purposes of error reduction and protection against possible (outlier) spurious readings. The performance of the multistage procedure specified in the draft EPA Recommended Practice was analyzed. Alternative methods of robust estimation published in the literature and developed for special application to the problem on hand

were reviewed and recommendations made which will improve performance. With respect to the CO procedure, specifically, it is recommended that substantial accuracy improvement can be achieved with no increase in numbers of FTP tests if the individual measurements used to establish the reference oil trend line are uniformly spread out over the 2000-mile interval instead of being clumped together as currently specified.

Section 7 presents a qualitative discussion of various formulations of fuel efficiency rating measures and the scaling assumptions implicit in them. For example, the form adopted in the draft procedure ratios the increase in mpg due to an oil to the base mpg. This suggests the belief that the same oil should provide twice the mpg increase in a car with twice the fuel economy (if ratings are to be invariant over models). The relationship of scaling assumptions to the basic mechanisms of friction reduction are discussed briefly.

3.0 FUEL EFFICIENCY RATING ERROR ANALYSIS

3.1 Non-Carryover Procedure

The most direct method of assessing the fuel efficiency performance of a candidate engine oil on a test vehicle would be to "age" the oil in that vehicle over a given mileage accumulation interval (specified at 2000 miles by EPA), measure fuel economy (F_C), flush the oil, replace with the high reference oil, measure fuel economy again (F_R), and ratio the two results. Thus, fuel efficiency rating of the candidate oil is expressed as

$$\varepsilon = \frac{F_C}{F_R}$$

with $\varepsilon > 1$ indicating some fuel efficiency effect (relative to the high reference oil) and $\varepsilon \leq 1$ indicating no effect or a relatively adverse effect of the oil. The validity of this method critically rests on the assumption that, following normal flushing of the candidate oil, there is no carryover of a residual efficiency effect due to the former oil. Hence, this procedure is referred to as the "non-carryover procedure."

As previously discussed in Falcon's "Analysis of Industry Comments," the variance of the error in ε from non-carryover (NCO) measurements on a single test vehicle is resolvable into three components, given approximately by:*

$$\sigma_\varepsilon^2 | \text{NCO} = \sigma_e^2 + \sigma_{oc}^2 + \sigma_{om}^2 .$$

σ_e^2 is the measurement error variance of a single fuel economy test result. Replicate tests are assumed carried out for each oil as specified

---

* Strictly speaking, batch-to-batch variability in oil characteristics, both candidate and reference, also contribute to total error. Industry responses however suggest this effect can be neglected.

in the EPA procedure or as modified for improved robustness against occasional spurious test results (see Section 6). In either case the variance per oil estimate is reduced to approximately one-half, but then combinations of both candidate and reference oil estimates doubles the total result back to σ_e^2. ASTM asserts in its comments to EPA that $\sigma_e \sim 0.75\%$ for volumetric or gravimetric techniques, while (by implication) $\sigma_e \sim 1.9\%$ for the carbon balance method of fuel economy measurement.

The second component σ_{oc}^2 refers to the variance in true fuel efficiency rating ε among sampled cars in the population of each test model (specified nameplate, model year, engine, etc.). Apparently, there are no direct experimental assessments of this variability, although the ASTM comments include an estimated upper bound of 1% for σ_{oc}.

The final component σ_{om}^2 refers to the variance in mean model fuel efficiency rating over the population of vehicle models. The last phrase needs to be made more definite. Ideally, this refers to all in-use models, weighted by their respective proportions in the fleet. In the interest of practicality, EPA has selected five specific models, equally weighted, to serve as proxies for the entire fleet. It is hence only feasible to measure variability over the five selected models, and σ_{om}^2 will then refer to this measure. It should be noted that this restriction can result in either underestimation or overestimation of the fleetwide σ_{om}^2, depending on the models selected. There exist no comprehensive data on the magnitude of σ_{om}. Union Oil Company included test results on a single oil in its comments which suggest that σ_{om} might be well below 0.25%. However, other respondents, notably Witco Chemical, suggest that large model-to-model differences do occur. The whole question of model-to-model variability, its impact on the significance of oil grading, and how one might deal with it are considered in more detail in Section 4.

If K cars of each of the five models are tested and the mean taken over the $5K$ individual fuel efficiency ratings, then the resulting error variance is reduced to:

$$\sigma_{\bar{\varepsilon}}^2 | NCO = \frac{\sigma_e^2 + \sigma_{oc}^2}{5K} + \frac{\sigma_{om}^2}{5}$$

Note that, in contrast to the first two error components, model variability error is not further reduced by replication over additional copies of the 5 designated models, i.e., increasing K. This is to be expected inasmuch as the same models continue to be tested. On the other hand, if σ_{om} is large, then as will be discussed in the next section one could question the meaningfulness of any single mean estimate of fuel efficiency estimate for the candidate oil.

In addition to the above described random errors, if a given candidate oil does have a carryover effect and is tested by the "back-to-back" non-carryover procedure, it will induce a systematic error, a shift in measurable fuel efficiency rating by amount $\Delta\varepsilon = -\varepsilon \cdot \Delta\tau$, where $0 \leq \Delta\tau \leq 1$. The limits indicate the possible range of carryover fraction: from full carryover, $\Delta\tau = 1$ (under which condition the car temporarily retains its candidate oil fuel economy level immediately after changeover to the reference oil), to zero carryover, $\Delta\tau = 0$. Fuel efficient oils are generally dichotomized on the basis of whether or not they exhibit a carryover effect. However, it is more likely that there is a continuum of carryover fractions.

For those oils with appreciable carryover fraction, the non-carryover procedure is clearly inadequate since no amount of replication will reduce the systematic error $\Delta\varepsilon$. A special flush procedure has been developed which is claimed to eliminate all or most of the carryover effect. If experimental investigations corroborate this claim, then, of course, all oils could be tested under the non-carryover procedure.

Concern has been expressed, however, that the neutralizing action of the special flush may be hard to control, leading to undercompensation in some cases and overcompensation in others--that is, the final level of fuel economy reached after replacement with reference oil could be higher or lower than would have been achieved had the candidate oil never been added.

3.2 Carryover Procedure

Under the supposition that there is no reliable way to eliminate carryover effects in some oils, and invoking the principle that the basic testing procedure should be the same for all oils, EPA has recommended adoption of a "carryover procedure." The basic concept is to measure reference oil fuel economy at several mileage points before aging of the candidate oil and linearly extrapolate to the candidate oil test mileage to provide a predicted reference oil fuel economy for comparison with candidate oil fuel economy. An underlying assumption of this procedure is that automobile fuel economy generally varies with accumulated mileage and that over relatively short intervals, say < 5000 miles, with cars that are sufficiently broken in (starting mileages > 10,000) this variation is sufficiently close to linear. Thus, as before, fuel efficiency rating is defined as a ratio:

$$\varepsilon = \frac{F_C}{F_R^*}$$

where F_R^* is now predicted instead of directly measured and consequently has additional error contributions. The variance of the error in this case (see Falcon's "Analysis of Industry Comments") is approximately:

$$\sigma_\varepsilon^2 | CO = \frac{\sigma_e^2}{2} + \sigma_{oc}^2 + \left(\sigma_m^2 + \frac{\sigma_e^2}{2}\right)\left(\frac{1}{n} + \frac{(X-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}\right) + \sigma_{om}^2$$

where: σ_m^2 = variance of unpredictable (small-scale) deviations in fuel economy from a linear trend as a function of mileage (under reference oil operation)

n = number of mileage points tested with reference oil

x_i = individual reference oil mileage points

$\bar{x} = \frac{1}{n} \Sigma x_i$

X = candidate oil mileage test point (at which the F_R^* prediction is made).

If we restrict consideration to the basic mileage interval configuration in the EPA procedure of 1000-1000-2000 corresponding to adoption of a trend line with three mileage points, then the prediction propagation factor takes on the numerical value of 4.83 and we can then write:

$$\sigma_\varepsilon^2|CO = 2.92\,\sigma_e^2 + \sigma_{oc}^2 + 4.83\,\sigma_m^2 + \sigma_{om}^2$$

3.3 Comparative Evaluation

Applying the previously noted estimates for σ_e and σ_{oc} together with an estimate of 2.34% for σ_m provided by ASTM, we find:

$$\sigma_\varepsilon^2|NCO = \begin{cases} (1.25\%)^2 + \sigma_{om}^2 & ; \text{ metered mpg} \\ (2.15\%)^2 + \sigma_{om}^2 & ; \text{ carbon balance mpg} \end{cases}$$

$$NCO \text{ bias} = -\varepsilon \cdot \Delta\tau$$

$$\sigma_\varepsilon^2|CO = \begin{cases} (5.39\%)^2 + \sigma_{om}^2 & ; \text{ metered mpg} \\ (6.16\%)^2 + \sigma_{om}^2 & ; \text{ carbon balance mpg} \end{cases}$$

Then, the corresponding errors in the mean estimates using 5K test vehicles are:

$$\sigma_{\bar{\varepsilon}}^2 | \text{NCO} = \begin{cases} \left(\frac{0.56}{\sqrt{K}}\ \%\right)^2 + 0.45\ \sigma_{om}^2 \\ \left(\frac{0.96}{\sqrt{K}}\ \%\right)^2 + 0.45\ \sigma_{om}^2 \end{cases}$$

$$\text{NCO bias} = -\bar{E} \cdot \Delta\tau$$

$$\sigma_{\bar{\varepsilon}}^2 | \text{CO} = \begin{cases} \left(\frac{2.41}{\sqrt{K}}\ \%\right)^2 + 0.45\ \sigma_{om}^2 \\ \left(\frac{2.75}{\sqrt{K}}\ \%\right)^2 + 0.45\ \sigma_{om}^2 \end{cases}$$

The σ_{om} term is kept separate since no estimate for it has been provided. However, one should keep in mind that this contribution should be small to justify a single grade determination for the candidate oil. Bearing in mind the postulated grade widths of the order of 2-3%, it is seen that the CO procedure, which was introduced to overcome the potential systematic error of $-\bar{\varepsilon} \cdot \Delta\tau$ in carryover oils, achieved this objective at the expense of a considerable increase in random error standard deviation. For example, it exhibits a magnitude comparable to grade width when $K=1$. To minimize classification errors in grading would require a fairly large K which, in turn, implies considerable testing expense and utilization of limited experimental resources.

The NCO procedure, on the other hand, may be able to yield acceptable grading accuracy for K's in the vicinity of 2 to 6, assuming of course, that the carryover effect is negligible or effectively eliminated by a special flush technique.

In the final analysis, it is essential to obtain reliable determinations of the various random error contributions and of the performance of special flushing techniques before reaching conclusions as to relative effectiveness or absolute feasibility of the CO and NCO procedures. The numerical estimates used in the foregoing analyses are admittedly tentative and are intended only to provide ballpark figures and suggest the likely direction of more accurately based conclusions.

4.0 MODEL-TO-MODEL VARIATIONS IN FUEL EFFICIENCY OF OILS

4.1 Grading Policies Under Model Differences

If a given candidate oil is accurately determined to have mean fuel efficiency ratings of 1.06 in model X cars, 1.02 in model Y cars and 0.97 in model Z cars, what grade should be assigned to that oil? Some industry comments suggest that this kind of a situation could very well occur, so it is far from a hypothetical question.

One policy approach might be to grade on the basis of a grand mean over all cars/models in the in-use fleet. The grade designation would then represent an average performance and the public would be educated to interpret it that way, allowing for the possibility that individual car performance could be considerably better or worse. An objection to this approach is that performance results on specific identifiable subsets of vehicles would be withheld from the public. If the oil in the above illustration did in fact receive an "average" grade of B, wouldn't the owners of X- and Z-model cars be misled by the label? A second objection is that averaging over the five selected test models may not necessarily provide a good estimate of the true fleet mean since the model sample is small and the criteria by which these test models were selected may have resulted in substantial bias. Considerable differences observed among the five tested models suggest comparable variability among the untested models with resulting high degree of uncertainty in the mean estimate.

An alternative policy might be to assign a grade based on worst model performance (or on the low end of the total range of uncertainty in calculated fuel efficiency ratings--including other error contributions). This in effect is what the EPA recommended procedure would accomplish with its LSD_{95} term. However, similar objections could be raised with respect to the withholding of useful information from the public and the possible unrepresentativeness of the five selected test models to the

total in-use fleet.

A further comment on the potential impact of the LSD_{95} term in the draft EPA Recommended Practice is perhaps in order. As noted above, it would tend to yield a conservative grade determination that is unlikely to exceed the poorest model-specific grade. An additional motivation for its inclusion, apparently, is to encourage the tester to achieve maximum feasible accuracy by control of experimental errors or by increased vehicle replication. If, however, the major contribution to the LSD_{95} term was to derive from between-model variability, then even substantial accuracy improvements would not appreciably reduce LSD_{95}, possibly much to the frustration of the tester.

A more complicated alternative, in those cases where it has been established that significant differences exist among the five test models, would be to assign multiple grades or a range of grades corresponding to the observed spread in performance. This approach partially meets some of the above objections, but the complexity of such labeling could conceivably cause increased confusion for some people.

Finally, it has been suggested that oils which exhibit significant differences among test models not be labeled at all. This approach, too, has its shortcomings. An oil manufacturer could devote considerable resources to a testing program which in the end yields no tangible result. Consider the case wherein a candidate oil tests grade B on all models except one on which it tests grade A. Such an oil would be left unlabeled despite the intuitive reasonableness of assigning it grade B. Another example is an oil that tests grade A on all models except one on which it tests grade B. The intuitive resolution of the problem in this case, however, is not as clear.

4.2 Statistical Test for Model Differences

We conclude that there is no completely satisfactory policy for grading oils that exhibit significant between-model variability. A policy will, of course, have to be established based on subjective evaluation of the various issues discussed above. In any event, one point is clear--it is important to determine whether or not significant between-model variability exists, and to estimate its magnitude. A proposed statistical procedure for accomplishing this is described in Appendix A. One conclusion stemming from the analysis in Appendix A is that the need for adequate statistical power to evaluate between-model variability will impose the severest requirement on number of vehicles to be tested. This aspect is discussed in more detail in Section 5.

4.3 Rationale for Selection of Test Models

The EPA recommended procedure is deficient in not indicating any rationale for actual model selection, other than the general phrase, "to represent a significant cross section of high volume production cars." This aspect of the procedure needs to be clarified because one should have a sound basis for generalizing from the tested models to the total in-use fleet, comprising mainly untested models.

The questions may be raised: why was it decided to test oils directly in automotive vehicles rather than by standard laboratory instrumentation, and then how was it determined that tests on five different types of vehicles would suffice? Apparently, an early decision was made that the realism and concrete interpretability of direct improvements in fuel economy observable in actual vehicles were important attributes that should not be given up in favor of more precise grading based on well-controlled laboratory measurement

of fundamental physico-chemical oil properties. On the other hand, practicality dictated that only a small sample of the very large number of operational vehicle types could be tested. One approach is to sample from the models randomly (with each model perhaps sales-weighted), but for sample sizes of the order of five, sampling errors can lead to highly unrepresentative selections. An alternative is to stratify the total population of models into a small number of distinct classes on the basis of some design parameter(s) known or believed to be predictive of fuel efficiency effects. Random sampling within each stratum could then assure a more representative mix of test models.

5.0 REQUIRED NUMBERS OF TEST VEHICLES

The overall objective of the procedures under consideration is assignment of a grade to each candidate engine oil tested that accurately characterizes its fuel efficiency effects in light-duty motor vehicles. An ancillary objective, as discussed in Section 4, is to determine whether or not there are significant differences in mean fuel efficiency rating among the five test models and, if so, to estimate model-specific efficiency ratings. In this section the number of test vehicles required to achieve these objectives at selected levels of confidence/precision will be analyzed for both the carryover and non-carryover procedures.

5.1 Grading Accuracy Requirements

We consider, first, achievement of accurate grading under the condition of model homogeneity, that is, no between-model variability in fuel efficiency rating. A reasonable negative measure of performance is the probability of incorrectly grading the oil. Drawing on an analysis previously reported in Falcon's "Analysis of Industry Comments," we note two alternative precise formulations of this measure:

P: the probability of misgrading an oil of a given true grade, assuming the fuel efficiency ratings of all oils of that grade to be uniformly distributed within the grade limits

Q: the probability of misgrading an oil with a true fuel efficiency rating at the center of its grade interval (most favorable case)

We shall restrict attention to only interior grades since the highest and lowest grades perform better. If an interior grade has width R and σ_ε is the standard deviation in estimated fuel efficiency rating, then it

is deduced from the previous analysis that:

$$P \tilde{=} 2\sqrt{\frac{2}{\pi}\frac{\sigma_{\bar{\varepsilon}}}{R}}$$

$$Q = 2\left[1 - \Phi\left(\frac{R}{2\sigma_{\bar{\varepsilon}}}\right)\right]$$

where Φ is the standard normal cumulative distribution function. Fixing R at 2.5% (the average of the 2% and 3% interior width in the EPA proposed grade structure), and utilizing the estimates for $\sigma_{\bar{\varepsilon}}|NCO$, $\sigma_{\bar{\varepsilon}}|CO$ derived in Section 3 (specifically selecting the more favorable experimental situation of metered mpg determination, assuming no carryover effect bias in the non-carryover procedure,* and setting $\sigma_{om} = 0$ consistent with our assumption of model homogeneity) we arrive at:

$$P_{NCO} \tilde{=} 0.36/\sqrt{K}$$

$$P_{CO} \tilde{=} 1.54/\sqrt{K} \qquad (K \geq 9)$$

$$Q_{NCO} = 2\left[1 - \Phi\left(2.23\sqrt{K}\right)\right]$$

$$Q_{CO} = 2\left[1 - \Phi\left(0.52\sqrt{K}\right)\right]$$

where K is the number of balanced model replications, i.e., total number of vehicles tested = 5K. Plots of P and Q as functions of K are shown in Figure 1. Imposition of a 10% average misgrading level, i.e., $P = 0.1$ yields $K = 12$ and ~210 for NCO and CO procedures, which implies 60 and over 1000 vehicles to be tested, respectively, per candidate oil. Note that satisfaction of $P = 0.1$ leads to vanishingly small Q. What this means is that misgrading probability is highly sensitive to the location of true fuel efficiency rating relative to grade boundaries and those oils closest contribute most of the errors.

---

* Carryover effect bias cannot be reduced by replication and should therefore not be considered in setting vehicle number requirements.

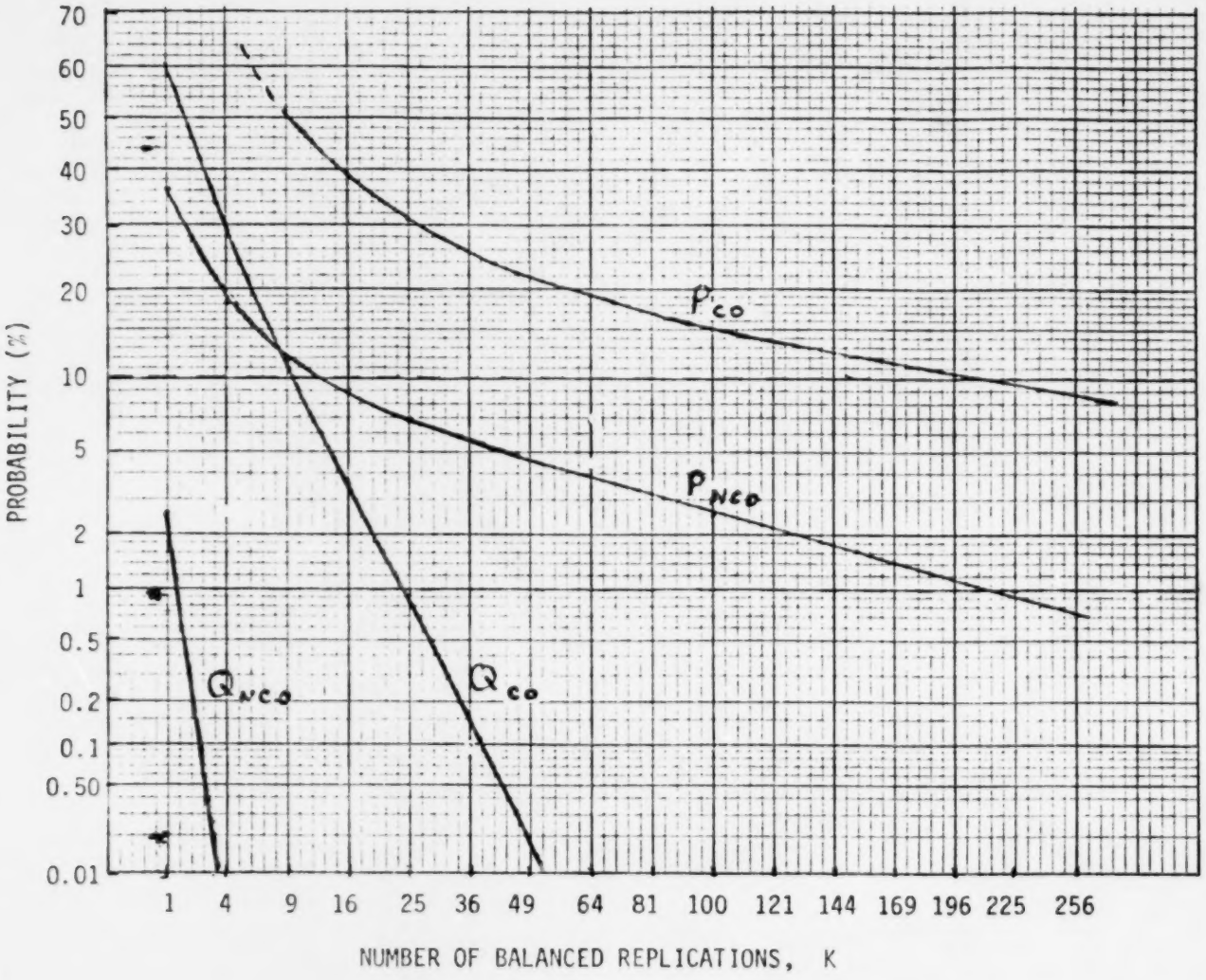

FIGURE 1. Misgrading Probabilities as a Function of Number of 5-Model Replications

If the 10% misgrading level criterion is accepted as reasonable and the various error component estimates given in the ASTM response are verified, then one would be forced to conclude that the carryover procedure requires an impractically large number of vehicles to be tested. Even the non-carryover procedure imposes a substantial requirement on numbers. Relaxation of the 10% misgrading level to 20% reduces the vehicle requirement for CO and NCO procedures to 20 and 300, respectively.

5.2 Statistical Test of Model Homogeneity

Now let us determine what requirements are imposed on vehicle numbers in order to test for significant model-to-model differences in fuel efficiency rating. Appendix A develops such a statistical test and establishes a precision parameter $\lambda = \sqrt{K/2}\,(R/\sigma_\varepsilon)$ which determines the power of the test for a given level of significance. Here R may again be taken as the (interior) grade width and σ_ε refers to the standard deviation in estimation of mean model fuel efficiency rating from a single test vehicle. Expressions for σ_ε are given in Section 3, but with the σ_{om}^2 term deleted since our focus here is on model-specific ratings.

We propose that our test have the following error properties: If the true maximum difference among the five model-specific fuel efficiency ratings exceeds grade width R, then we fail to declare the models as significantly different in at most 10% of such cases; conversely, if the true maximum difference among the five ratings is less than $0.75\,R$, then we fail to declare the models as effectively equivalent in at most 10% of such cases. It is shown that this performance is achievable with $\lambda = 10$. If the above performance specification is relaxed by substituting $0.50\,R$ in place of $0.75\,R$, then the corresponding λ is reduced to 5. Other error or statistical power properties can

be specified, but it is believed the above two examples encompass reasonable minimum requirements. What are the implications for number of test vehicles? Again using ASTM error estimates with metered mpg experimental setup, as provided in Section 3, we deduce:

$$K = 2\lambda^2 \frac{\sigma_{\bar{\varepsilon}}^2}{R^2} = \begin{cases} 50; & \text{NCO}, \ \lambda = 10 \\ 930; & \text{CO}, \ \lambda = 10 \\ 13; & \text{NCO}, \ \lambda = 5 \\ 232; & \text{CO}, \ \lambda = 5 \end{cases}$$

Hence for total test vehicles

$$5K = \begin{cases} 250; & \text{NCO}, \ \lambda = 10 \\ 4650; & \text{CO}, \ \lambda = 10 \\ 65; & \text{NCO}, \ \lambda = 5 \\ 1660; & \text{CO}, \ \lambda = 5 \end{cases}$$

Comparison with requirements based on control of misgrading errors shows the present requirements to dominate. In particular, the CO procedure is basically impractical if one phase of the statistical data analysis requires separate determination for each candidate oil as to whether or not its fuel efficiency effects are substantively different among the models tested. The NCO procedure could be construed, perhaps, as marginally feasible depending on tradeoff of testing costs with incremental value to the oil manufacturer of establishing a favorable grade label.

What is the basic problem? The analysis makes it clear that measurement/sampling errors must be roughly an order of magnitude smaller than the oil effect differences to be resolved. The inherent precision of fuel economy measurement and uniformity among copies of the same vehicle

model class are simply not sharp enough, and this lack of precision can only be overcome by a very large number of replications.

Finally, we must emphasize that the conclusions reached here are based on error component estimates that are not all firmly established. Substantial reduction, for example, in the estimate for fuel economy variability under extended mileage accumulation σ_m^2 could put the CO procedure in a very much more favorable light. Also, if an experimental study involving carefully selected samples of "fuel efficient" oil brands and vehicle models were to establish the absence of any substantive between-model variabilities, then there would be no necessity for such a statistical test to be performed as part of the procedure for each candidate oil.

6.0 STATISTICAL STRATEGIES IN FUEL ECONOMY MEASUREMENT

6.1 Non-Carryover Procedure

6.1.1 Error Reduction Through Replication

We first consider the non-carryover procedure as described in Section 2. No mileage accumulation effect and extrapolation errors are involved. Assume no distinction among test models so that the between-model variability issue does not arise. If a single fuel economy measurement is made with candidate oil in a sampled vehicle and a single measurement again after replacement with reference oil, then the error variance of the ratio (fuel efficiency ratios) is approximately

$$\sigma_\varepsilon^2 = 2\,\sigma_e^2 + \sigma_{oc}^2$$

where σ_e and σ_{oc} are the standard deviations of test replication error and car-to-car differences in true rating, respectively. ASTM estimates $\sigma_e = 0.75\%$ and $\sigma_{oe} = 1\%$. It then makes sense to achieve some reduction in total error variance by retesting the same vehicle which is a relatively inexpensive means of replication. Thus, with n replicate tests

$$\sigma_\varepsilon^2 = 2\,\sigma_e^2/n + \sigma_{oc}^2$$

On the other hand it is also clear (assuming ASTM's error estimates) that relatively little will be gained for the effort expended with n beyond 2 or 3 because of the increasing dominance of σ_{oc}^2. Further replication is best done by sampling more vehicles. An additional motivation for test replication on the same vehicle is to provide protection against spurious observations, and this purpose, too, can be accomplished to some level of confidence with $n = 2$ or 3.

EPA has in fact incorporated test replication at variable levels of $n = 2, 3, \text{or } 4$ in its recommended practice. Admittedly, it was done in the context of the carryover procedure involving trend line extrapolation, but the basic principles of test error reduction and protection against spurious readings still apply.

6.1.2 Robust Estimation

Being troubled by the apparent inefficiency of the EPA procedure, which bears some resemblance to a "best two out of three" rule (see our discussion on this point in Falcon's "Analysis of Industry Comments" and Appendix C), we decided to look more deeply into the problem of small sample robust estimation. The resulting analysis is included as Appendix B. Also, some earlier simulation results on performance of the EPA procedure, previously alluded to but not explicitly documented, are provided in Appendix D.

Among various estimators considered in Appendix B, two which were determined to have particularly good robustness properties in the face of possible spurious contamination of a single reading are the Veale-Huntsberger estimator based on three replicate measurements and the K-estimator which is two-staged and may take either two or three observations. The Veale-Huntsberger estimator is a weighted mean of three observations in which the weight corresponding to the observation most deviant from the unweighted sample mean decreases gradually with increasing deviation until a criterion deviation (in standard deviation units) is reached, beyond which the assigned weight drops to zero, i.e., the deviant observation is totally rejected. The K-estimator takes the mean of the first two observations if they are closer than a preset criterion value (in standard deviation units); otherwise it requires a third observation to be made and then reverts to a Veale-Huntsberger estimator.

If the non-carryover procedure can accept the cost of three back-to-back FTP measurements per fuel economy determination, then the Veale-Huntsberger estimator is recommended. For a particular criterion value (c=2.4042), the root mean square error of estimation (RMSE) under conditions of no contamination and worst case single spurious reading are $0.589\ \sigma_e$ and $1.091\ \sigma_e$, respectively, where σ_e is the single-measurement standard deviation. This compares with the best possible RMSE of $\sigma_e/\sqrt{3} = 0.577\ \sigma_e$ achieved by the sample mean only under no contamination. The penalty paid under no contamination for using this estimator rather than the sample mean is $(0.589 - 0.577)/0.577 = 2\%$ increase in standard deviation. On the other hand, the gain in performance under contamination is dramatic since the RMSE of the sample mean increases without bound as a function of the magnitude of the contamination.

If, on the other hand, measurement costs make it prudent to minimize the number of replications, while at the same time maintaining some robustness property, then the K-estimator is recommended as having superior performance to the method proposed in the draft EPA Recommended Practice. Although further exploration of the K-estimator as a function of criterion values is needed for detailed optimization, an indication of its capability is given by simulation results for selected criteria, shown in Appendix B. In one particular case ($c = d = 2$), the mean number of replications per fuel economy determination is bounded by 2.22 (under the assumption of 0.05 probability of contamination), while the RMSE achieved under conditions of no contamination and worst case single spurious reading are $0.704\ \sigma_e$ and $1.078\ \sigma_e$, respectively. This is superior to the procedure in the draft EPA Recommended Practice (closely approximated by the E-estimator in the Appendix with $d = 2$) which takes somewhat fewer observations on the average (bounded by 2.06) and yields no-contamination RMSE and worst-case RMSE of $0.723\ \sigma_e$ and $1.180\ \sigma_e$, repsectively. The noted K-estimator also compares favorably with the

best possible RMSE (for N=2.22) of 0.681 σ_e achieved by a random combination of 2-sample and 3-sample means under no contamination. The penalties paid under no contamination, in terms of percent increase in error standard deviation, are hence: (.704 - .681)/.681 = 3.4% for the K-estimator and (.723 - .681)/.681 = 6.2% for the EPA procedure.

6.2 Carryover Procedure

6.2.1 Dominance of Extended Mileage Accumulation Variability

We next consider the carryover procedure as described in Section 2, which is the procedure incorporated in the draft EPA Recommended Practice. Again, based on unreplicated* fuel economy testing on a single sampled vehicle (and between-model variability assumed zero), the error variance of the calculated fuel efficiency rating would be approximately (see Section 3):

$$\sigma_\varepsilon^2 = \sigma_e^2 + \sigma_{oc}^2 + 4.83\,(\sigma_m^2 + \sigma_e^2)$$

$$= 5.83\,\sigma_e^2 + 4.83\,\sigma_m^2 + \sigma_{oc}^2$$

This result involves three single reference oil mpg determinations separated by successive 1000-mile intervals and a single candidate oil mpg determination following a 2000-mile aging interval. Introducing the component error estimates by ASTM we may write:

$$\sigma_\varepsilon^2 = 5.83\,(0.56) + 4.83\,(5.48) + 1.0 = 30.7 \quad (\%)^2$$

$$\sigma_\varepsilon = 5.54\%$$

---

* That is, wherever an FTP fuel economy measurement is called for it is done only once instead of at least twice, back-to-back, with possible third or fourth measurements, as called for in the draft Recommended Practice.

We observe at once that the dominant error contribution is due to σ_m^2, which, it will be recalled, is the variance of the deviation of fuel economy of a car from a linear trend under extended mileage accumulation. If fact, if measurement error itself were totally eliminated, σ_ε would decrease only slightly to 5.24%!

6.2.2 Proposed Alternative Experimental Design

EPA, as previously described in its draft Recommended Practice, requires replicated mpg determinations at each of the test mileage points. Clearly this would yield little improvement in accuracy under normal circumstances. The sole justification for replication in this situation must therefore be to achieve robustness in the event of spurious observations. That would be accomplished to a significant degree as indicated in our discussion of the non-carryover procedure. However, as long as additional FTP tests are to be made, we propose a different placement of these tests in order to achieve a more substantial gain in precision along with the protection against spurious observations. Specifically we suggest that the six FTP tests with the reference oil (which is approximately the expected number of such tests under the presently formulated procedure) be performed singly at 400 mile spacings to cover the same total range of 2000 miles. It is believed that the cost impact of such a modification would be relatively small. The FTP test with the candidate oil could continue to be replicated in the present manner or could apply a K-estimator as discussed earlier. Under this modified design, the error variance of ε would reduce approximately to:

$$\sigma_\varepsilon^2 = \frac{\sigma_e^2}{2} + \sigma_{oc}^2 + 3.38\,(\sigma_m^2 + \sigma_e^2)$$

$$= 3.88\,\sigma_e^2 + 3.38\,\sigma_m^2 + \sigma_{oe}^2$$

$$= 21.7\ (\%)^2$$

$$\sigma_\varepsilon = 4.66\%$$

The major reduction comes from the reduced extrapolation coefficient of the dominant σ_m^2 term. This result assumes that the frequency spectrum of unpredictable mpg variations with mileage is mostly in a high enough region so that variations separated by 400 mile increments remain essentially uncorrelated. If appreciable "low frequency components" do exist then the accuracy gain would not be as great. This issue needs to be experimentally resolved.

Accurate estimation in the face of possible spurious observations can continue to be achieved through robust procedures for linear regression. Detailed discussion of specific methods is not herein provided, since an adequate literature on this subject exists.

It is interesting to observe that the coefficient of the σ_e^2 term, 3.88, in the proposed evenly spaced design is in fact larger than the comparable coefficient, 2.92, for the present design of three paired measurements separated by successive 1000-mile increments (see Section 3). This is explained by the fact that measurement errors are uncorrelated in either design and the present EPA design, by its more extreme placement of tests, yields lower extrapolation error due to that component. In the case of the mileage accumulation variation component, the decorrelation achieved in the modified design more than makes up for its reduced extrapolation capability.

6.3 Concluding Remarks

In conclusion, if EPA were to finalize a recommended practice based on a non-carryover procedure, we recommend consideration of either the Veale-Huntsberger or K-estimators for robust and reasonably efficient estimation of FTP fuel economy utilizing small numbers (≤ 3) of replicated

The major reduction comes from the reduced extrapolation coefficient of the dominant σ_m^2 term. This result assumes that the frequency spectrum of unpredictable mpg variations with mileage is mostly in a high enough region so that variations separated by 400 mile increments remain essentially uncorrelated. If appreciable "low frequency components" do exist then the accuracy gain would not be as great. This issue needs to be experimentally resolved.

Accurate estimation in the face of possible spurious observations can continue to be achieved through robust procedures for linear regression. Detailed discussion of specific methods is not herein provided, since an adequate literature on this subject exists.

It is interesting to observe that the coefficient of the σ_e^2 term, 3.88, in the proposed evenly spaced design is in fact larger than the comparable coefficient, 2.92, for the present design of three paired measurements separated by successive 1000-mile increments (see Section 3). This is explained by the fact that measurement errors are uncorrelated in either design and the present EPA design, by its more extreme placement of tests, yields lower extrapolation error due to that component. In the case of the mileage accumulation variation component, the decorrelation achieved in the modified design more than makes up for its reduced extrapolation capability.

6.3 Concluding Remarks

In conclusion, if EPA were to finalize a recommended practice based on a non-carryover procedure, we recommend consideration of either the Veale-Huntsberger or K-estimators for robust and reasonably efficient estimation of FTP fuel economy utilizing small numbers (≤ 3) of replicated

7.0 SCALING ASSUMPTIONS IMPLICIT IN SPECIFIC FUEL EFFICIENCY RATING MEASURES

The definition of "grades" for candidate fuel-efficient oils must be based on some measure of "performance" of these oils in the sense of fuel saving. What constitutes an appropriate "figure of merit" is a basic question which needs to be addressed before any statistical procedure is invoked for the establishing of grades.

This section of the report contains some observations pertinent to the parameterization of the "benefit" to be derived from an alleged fuel-efficient oil. It considers both physical and statistical implications of the choice of a performance measure and concludes with some further observations pertinent to the LSD_{95} aspects of the EPA Recommended Standard Practice.

The performance measure proposed in the EPA Recommended Practice is based, in essence, on percentage improvement in fuel economy as determined by comparative tests of candidate oil (CO) and high reference oil (HR) in five species of vehicles. Specifically, one computes

$$\varepsilon = \frac{C - LSD_{95}}{H}$$

where C is the mean fuel economy as experienced with the candidate oil and H is the mean fuel economy as experienced with the reference oil. Fuel economy has its usual connotation of miles per gallon (MPG), and the means are computed in MPG space for K vehicles of each of the five species. Were it not for the LSD_{95} term, the performance measure would translate into a simple ratio of candidate-oil mean MPG to reference-oil mean MPG.

Questions pertaining to the LSD_{95} aspects of the measure are discussed later in this section. Here we are concerned with the more basic question of parameterization of "improvement" in some sense. We assert, ipso facto, that other bases of comparison of the two oils exist and that the use of ε, as defined in the EPA procedure, needs to be defended on more than its intuitive appeal. In particular, the notion of ε seems to impose an arbitrary "scaling law" which may not be realistic. It implies that the more fuel efficient a car is, the more will its fuel economy (MPG) be improved by the use of the fuel-efficient oil.

The realism of such an assumption is by no means obvious, nor is it necessarily consistent with physical theories of lubrication and the friction process. Though it might be argued that energy losses due to friction are proportional to distance traveled, this argument does not necessarily hold in the aggregate of all vehicles. Different regimes of lubrication are known to exist in different automobiles, and the greater fuel economy of a particularly fuel-efficient vehicle could stem from causes unrelated to its lubrication characteristics. The assumption that the benefit to be derived from a fuel efficient oil is proportional to the base-level fuel efficiency of the vehicle may therefore be untenable.

The importance of scaling can be appreciated by examining its effect on the computations involved in the determination of ε. Among the test vehicles, observed improvements in fuel economy can be expected to vary. These variations are due, in part, to errors of sampling and errors of measurement. Another possible source of variation, however, stems from the fact that the mean response of vehicles to the candidate oil may be vehicle-species dependent. If this species-related variation is in accordance with the proportional scaling implied in the EPA Recommended

Standard Practice, then the expected relative improvement in fuel efficiency attributable to the candidate oil is constant across vehicle species, even though the differential improvement in MPG is not. Under such conditions, the choice of vehicle species to be included in the test sample could be arbitrary, because the mean relative improvement in fuel economy would be independent of species selection.

If the scaling assumption does not hold, then the value of ε is an artifact of the choice of vehicle species to be included in the test. This fact would not preclude the possibility, however, of some other scaling law and an associated figure of merit. What one would hope for is some basis of comparison which would be invariant with regard to vehicle species--in short, one hopes to find a "scaling parameter" applicable to all vehicles. Undoubtedly, absolute invariance does not exist, but the figure of merit should be "robust"--that is, as insensitive to vehicle characteristics as possible.

One might well ask whether evaluation of fuel efficiency improvement should be based on fuel consumption (gallons per mile) rather than MPG. For small perturbations, of the order of a few percent, the bases do not differ substantially. In other words, a 5% increase in miles per gallon translates to good approximation into a 5% reduction in fuel consumption. The gallons-per-mile point of view is not without merit, however, in view of the fact that the matter of concern is fuel conservation for a given distance traveled. For a test based on several vehicle species with a range of fuel economy ratings, the MPG and GPM bases could still lead to different evaluations. An illustration of this fact follows.

Let $R_1, R_2, \ldots R_5$ denote the reference-oil fuel economies in miles per gallon of the five vehicle species tested. Similarly, let k_i denote the factor of improvement which scales the reference MPG for the ith species into the fuel economy C_i experienced with the candidate oil. Thus,

$$C_i = k_i R_i$$

and we can define a quantity ε' as

$$\varepsilon' = \frac{\Sigma k_i R_i}{\Sigma R_i}$$

Note that ε' is analogous to the quantity ε defined in the EPA Recommended Practice except for the deletion of the LSD_{95} term.

Now if $k_i = k$ for all i, then $\varepsilon' = k$, where k is expected to have a value quite close to unity.

But, suppose that the k_i assume distinct values which vary over an appreciable range for the five species. Then it is clear that the R_i serve as weighting factors and that most weight is given to those vehicles which normally achieve the highest MPG. For example, if $R_1 = 15$ MPG and $R_5 = 30$ MPG, then k_5 is given twice as much weight as k_1.

Now one can envision various scenarios, two of the most interesting being (a) the case in which k values monotonically increase with MPG and (b) the case in which k values monotonically decrease with MPG. In case (a), the candidate oil would receive a relatively high ε' because of the heavy weighting accorded the high-MPG vehicles. In case (b), the candidate oil would receive a relatively low ε' because of the

heavy weighting accorded the low-MPG vehicles. It is a moot point, however, as to which value would most realistically represent the fuel efficiency of the oil, in view of the fact that the low-MPG cars have greatest fuel consumption. Accordingly, there may be merit in comparing candidate and reference oils on the basis of consumption, in GPM, rather than on the basis of MPG.

The question of what constitutes proper parameterization of fuel-efficiency improvement is made more complex by introduction of the LSD_{95} term. From a statistical point of view, the use of the LSD_{95} term is illogical, particularly so if the true k_i differ significantly from species to species. If the observed values vary only because of measurement error, then each observation is an estimate of the common, population value k for the candidate oil in question. Only then can averaging and other statistical procedures be defended on the basis of improved estimation of ε. Even in that case, however, the proposed use of LSD_{95} confuses population parameters with estimates based on realizations of random variables. This point needs further explication.

The procedure recommended by EPA comes under the heading of what is often referred to in the statistical literature as a paired comparison. Two random variables X and Y with realizations x_i and y_i $(i = 1, 2, \ldots, N)$ are arrayed as in the following table.

A PAIRED COMPARISON

VARIABLE	X	Y	DIFFERENCE, D
OBSERVATIONS	x_1	y_1	$d_1 = x_1 - y_1$
	x_2	y_2	$d_2 = x_2 - y_2$
	⋮	⋮	⋮
	x_n	y_n	$d_n = x_n - y_n$
MEAN	$\bar{x}$	$\bar{y}$	$\bar{d} = \bar{x} - \bar{y}$

One can then calculate the standard error of the d_i and test the hypothesis that the true mean difference m_D is zero. The level of significance can be as desired, and one can choose either a one-sided or two-sided alternative hypothesis. If the 95% level is selected, then the LSD_{95} is simply the value of $\bar{d}$ required to reject the hypothesis H_0: $m_D = 0$.

The important point to note here is that the LSD_{95} is a sample-derived quantity which serves as a criterion for rejection of the null hypothesis. The EPA Recommended Practice proceeds to compute the amount by which $\bar{d}$ exceeds this criterion and to use this excess as the basis for defining oil grade. It is our belief, however, that the intent of the procedure is not served by adjusting the observed $\bar{d}$ by a sample quantity. Rather, it should be adjusted by a fixed quantity, that fixed quantity being one of the grade boundaries.

For example, suppose that the mean difference $\bar{\Delta}$ between the candidate oil and the reference oil is 1 mpg with a standard error of 0.3 mpg. Suppose, further, that the sample mean fuel economy for the reference oil is $H = 20$ mpg. For 4 degrees of freedom, $t = 2.1318$ and

$$\varepsilon = \frac{21.0 - 2.1318(0.30)}{20} = 1.018.$$

Accordingly, by the EPA Recommended Practice, the oil would qualify as Grade B.

Note that the use of the LSD_{95} procedure is equivalent to computation of a one-sided confidence interval for the population mean. The lower confidence bound is provided by

$$21.0 - 2.132(0.3) = 20.36,$$

or, in ratio terms relative to $H = 20.0$, the 1.018 seen above. In other words, the statement that the true mean exceeds 1.018 has 95% probability of being a correct statement. It might well be asked if there is not a "reasonably high" probability that the true mean exceeds 1.03.

This prospect can be examined as follows. Suppose the true mean is 1.03. The computation

$$\frac{21.0 - 1.03(20)}{0.3} = 1.333$$

provides a t value corresponding to nearly a 90% confidence interval

($t_{0.1} = 1.533$). Similarly, if the true mean is 1.01, the computation

$$\frac{21.0 - 1.01\ (20)}{0.3} = 2.666$$

provides a t value corresponding to nearly a 97.5% confidence interval ($t_{0.025} = 2.776$). In other words, it is "almost as likely" that the true mean exceeds 1.03 as that it exceeds 1.018 or even 1.01. Indeed, it can even be argued that the grade class is more likely to be A than B.

The difficulty is further illuminated when viewed in a different light. The LSD_{95} procedure, as employed by EPA, addresses the question of bounding the true mean improvement <u>given</u> a sample measurement. An equally important question concerns bounding the sample measurement, <u>given</u> the true mean improvement. For example, suppose that the true value for a candidate oil is really (1.01)(20) = 20.2--that is, at the low end of the B-grade interval. With what probability would such oils yield a sample value exceeding the LSD_{95} criterion? Clearly, one is concerned with the distribution of the quantity

$$\frac{x - 20.2}{s}$$

where x is the observed value of the sample mean for the test vehicles and s is the estimated standard error. By a previous argument, the event

$$\frac{x - 20.2}{s} > 2.132$$

would occur with probability 0.05, so that one would seldom "recognize" this oil to be a B-grade oil by the EPA procedure.

Admittedly $\varepsilon = 1.01$ barely qualifies the oil as Grade B, but a minimal A-grade oil, i.e., an oil with $\varepsilon > 1.03$, would fare only slightly better, as will be shown by the following argument. With $\varepsilon = 1.03$ and reference value of 20 MPG, the expected value for the candidate oil is $(1.03)(20) = 20.6$ MPG. Therefore each observed value would be 0.4 MPG greater than in the previously postulated case in which $\varepsilon = 1.01$ was assumed. Accordingly, one is concerned with the distribution of the quantity

$$\frac{(X + 0.4) - 20.2}{s} = \frac{(X - 20.2) + 0.4}{s}$$

and with the probability of occurrence of the event

$$\frac{(x - 20.2) + 0.4}{s} > 2.132.$$

A bit of algebra shows that the probability of this event is the same as that of the event

$$\frac{x - 20.2}{s} > 0.799$$

By reference to tables of the t distribution, it is evident that a Grade B assignment or better would result "less than 25% of the time" ($t_{0.25} = 0.741$) even if the true improvement were at the upper limit of the B range.

In summary, we have examined the procedure whereby the EPA Recommended Standard Practice combines measurements from several test vehicles into a single oil grade rating. It is concluded that, from a physical point of view, the procedure is tantamount to an arbitrarily assumed scaling law when

the vehicles tested differ significantly in their response to the fuel-efficient oil. It is further concluded that the use of statistical concepts such as the LSD_{95} is appropriate only if the observed differences among vehicles are due to sampling and measurement errors.

When examined in this light, it is concluded that the LSD_{95} procedure, as employed in the EPA recommended standard practice, is structured so as to guard against overgrading a candidate oil. On the other hand, when viewed from the standpoint of both Type I and Type II errors, it tends to undergrade oils, perhaps seriously. It must be concluded, therefore, that the LSD_{95} approach does not afford the supposed 95% level of significance in the sense desired and should be replaced by a hierarchy or sequence of hypothesis tests involving the postulated grade boundaries. These tests should be constructed in such a way as to balance the two types of error, overgrading and undergrading.

APPENDIX A

ENGINE OIL FUEL EFFICIENCY GRADING UNDER MEASUREMENT ERROR AND BETWEEN-MODEL DIFFERENCES

If C and H are the true FTP fuel economies achieved by a vehicle with candidate oil and reference oil, respectively, then $\varepsilon = C/H$ is defined as the candidate oil's fuel efficiency rating with respect to the given vehicle.

Let $G_1 < G_2 < \ldots < G_r$ be a grading structure such that an oil is designated as Grade A if $\varepsilon > G_r$, Grade B if $G_{r-1} < \varepsilon \leq G_r$, and so forth. This grading structure will be characterized by a minimum grade width, $R = G_j - G_{j-1}$ for some $j = 2, \ldots, r$.

The EPA proposed procedure for grading a candidate oil involves testing the oil in a balanced sample from five preselected vehicle models. Specifically, designate ε_{ik} as the oil's measured fuel efficiency rating in the k^{th} car of the i^{th} model, where $i = 1, \ldots, 5$ and $k = 1, \ldots, K$. The basic issue addressed here is: under what conditions of measurement error, car-to-car variability, and model-to-model differences can a reasonable statistical procedure be applied to the data (ε_{ik}) to obtain an accurate and meaningful fuel efficiency grade? A subsidiary question, of

course, is what constraints need to be imposed on K, the number of test cars per model?

We assume an error/effects model for the data as follows:

$$\varepsilon_{ik} = \mu_i + e_{ik}$$

where μ_i is the true mean value of candidate oil fuel efficiency rating for cars of the model i population, and e_{ik} is an additive error term which includes both experimental (measurement) errors and sampling effects (vehicle-to-vehicle variability within model i). It is also believed reasonable to assume that the e_{ik} are independent, normally distributed variates with zero mean and common variance σ^2. One may, further, introduce an external estimate S_o^2 of σ^2 based on prior experience/data. Previous analysis by ASTM[1] provides estimates for experimental error standard deviation of 1.6% under the non-carryover procedure and 6% under the carryover procedure. These are therefore lower bounds for S_o, but if it could be established that sampling effects are relatively small, then they would serve directly as appropriate values of S_o for the two distinct experimental procedures.

Between-model variability is represented in the above formulation by the individual model means. Unfortunately, we know very little about the

Reproduced from best available copy.

magnitude of this effect at the present time, yet it could conceivably negate the basic validity of the grading program. If model-to-model differences for a given oil are large compared to grade width measure R, then one could quite justifiably question the meaningfulness of a specific grade assigned to that oil. Our recommendation in such an event is that it is better to leave the oil ungraded than to assign a fictitious "average" or "worst" grade. An "average" grade could give the owner of one of the poorly performing models a false sense of value. A "worst" grade is likely to be so low that the oil company would rather opt for no grade.

For purposes of the present analysis we regard the five models pre-selected by EPA as the universe of models. The rationale for the selections made and their representativeness are clearly relevant, but will not be pursued here.

The basic approach which we propose is to test the hypothesis that between-model differences are substantial relative to grade width. If the data permit us to reject this hypothesis, then a fortiori the measurement/sampling error contribution is acceptably small and we would then estimate the oil's fuel efficiency rating by the total sample mean:

$$\bar{\varepsilon} = \frac{1}{5K} \sum_{k=1}^{K} \sum_{i=1}^{5} \varepsilon_{ik}$$

We would then proceed to determine the grade interval which contains $\bar{\varepsilon}$ and assign the oil that grade.

If, conversely, the data cause us to accept the hypothesis, that is equivalent to saying we have insufficient confidence that the between-model variability is small enough to permit a meaningful grade to be assigned. The oil company (tester) may decide at this point to take additional measurements (increase K) and repeat the above test with the augmented data. It could also decide not to continue in which case the oil may be left ungraded or some prescribed algorithm based on estimated model-specific ratings may be applied to determine a single or multiple grade outcome. Testing costs will, of course, put practical constraints on K. However, consideration should also be given to the advisability of EPA imposing a maximum limit on K.

A Proposed Test for Between-Model Variability

The form of the hypothesis we wish to test is

$$H_o: \max_{i,j} \{|\mu_i - \mu_j|\} \geq \rho R$$

where $\rho > 0$ is a fixed number which establishes the severity of the test. Define $\Delta\mu = \max \{|\mu_i - \mu_j|\}$ and $R' = \rho R$ and restate the above as $H_o: \Delta\mu \geq R'$.

Reproduced from best available copy.

The hypothesis states that there exists a pair of car models whose difference in true mean fuel efficiency rating is at least ρ times the smallest grade width. In other words, H_o affirms that model-to-model variability in mean fuel efficiency rating is too large relative to grade width to permit meaningful determination of a specific grade. A reasonable value for ρ is 1, in which case rejection of H_o would imply that the model variability contribution to rating error is likely to be less than half a grade width. Other ρ values could be specified. As will be seen later, specification of ρ and of level of significance are interchangeable decisions.

What is a reasonable statistic for testing this hypothesis? Define the within-model sample means $\bar{\varepsilon}_i$ (estimates of the μ_i) by:

$$\bar{\varepsilon}_i = \frac{1}{K} \sum_{k=1}^{K} \varepsilon_{ik} \quad i = 1, \ldots, 5$$

The common variance of these sample means is σ^2/K. Some alternative estimates of σ^2 are:

$$\hat{\sigma}^2 = S_o^2$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{5\,(K-1)} \sum_{i=1}^{5} \sum_{k=1}^{K} (\varepsilon_{ik} - \bar{\varepsilon}_i)^2 \; ; \; K \geq 2$$

$$\hat{\sigma}^2 = \frac{\nu S_0^2 + 5(K-1)\,S^2}{\nu + 5(K-1)} \quad ; K \geq 2$$

where S_0^2 as previously noted is an external estimate of σ^2 and is assumed to be chi-square with ν degrees of freedom. Note that "internal" estimation of σ^2 is possible only for $K \geq 2$, i.e., some within-model replication is required. The third alternative, which uses all available information, is clearly best under the assumption that σ^2 has not changed. For purposes of the present analyses we shall assume that an S_0^2 is available with sufficiently large ν to be effectively equivalent to known σ^2.

Now, let $r_1 \leq r_2 \leq \ldots, \leq r_5$ be the order statistic of $\{\bar{\varepsilon}_i\}$. Define the statistic

$$T = \sqrt{\frac{K}{2}} \cdot \frac{r_5 - r_1 - R'}{\sigma}$$

Under the assumption of relatively precise estimation ($\sigma << R'$) and neglectability of intermediate μ_i,* T will be approximately normal with mean equal to $\sqrt{K/2}\,(\Delta\mu - R')/\sigma$ and unity variance. Consequently, we

---

* That is, there exists a unique pair of extreme μ_i and the other three μ_i have intermediate values which are relatively far from the extremes in σ units.

may make the following probability statement:

$$\Pr\{T < z_{1-\alpha} \mid H_o\} \leq \alpha$$

Therefore, an acceptance region for H_o defined by $T \geq -z_{1-\alpha}$ constitutes a test with level of significance $\leq \alpha$ under the above stated assumptions. It is interesting to observe that if, contrary to the assumption, more than two μ_i are extreme or close to extreme, then T will be biased in a positive direction and we will be more apt to accept H_o (which affirms that model-to-model differences are too great). This effect nicely matches our inclination since for fixed $\Delta\mu$ polarization of the μ_i corresponds to an increase in variability.

Power of the Test

Suppose that $\Delta\mu$ is truly smaller than the criterion value of R'. What is the probability of making the correct decision to reject H_o? This is the power, P, of the test and will, of course, be a function of $\Delta\mu$. Again, invoke the previously stated assumptions of $\sigma << R'$ and neglectability of intermediate μ. Define:

$$T' = T - \sqrt{\frac{K}{2}}\,(\Delta\mu - R')/\sigma$$

which is approximately standard normal. We may then write:

$$P = Pr\{T < -z_{1-\alpha} | \Delta\mu\}$$

$$= Pr\{T' < -z_{1-\alpha} \sqrt{\frac{K}{2}} (\Delta\mu - R')/\sigma\}$$

$$= \Phi(-z_{1-\alpha} \sqrt{\frac{K}{2}} (\Delta\mu - R')/\sigma)$$

where Φ is the normal cumulative distribution function. In terms of the dimensionless parameters,

$$\delta = \frac{\Delta\mu}{R'} \quad , \quad \lambda = \sqrt{\frac{K}{2}} \frac{R'}{\sigma}$$

this expression becomes

$$P = \Phi(-z_{1-\alpha} + \lambda(1 - \delta))$$

Figure 1 shown plots of P vs. δ for various values of significance level α and precision parameter λ. We focus on the interval $0 \leq \delta < 1$ which corresponds to the alternate hypothesis $H_1: \Delta\mu < R'$. At $\delta = 1$, power equals level of significance, but increases with decreasing δ.

Reproduced from best available copy.

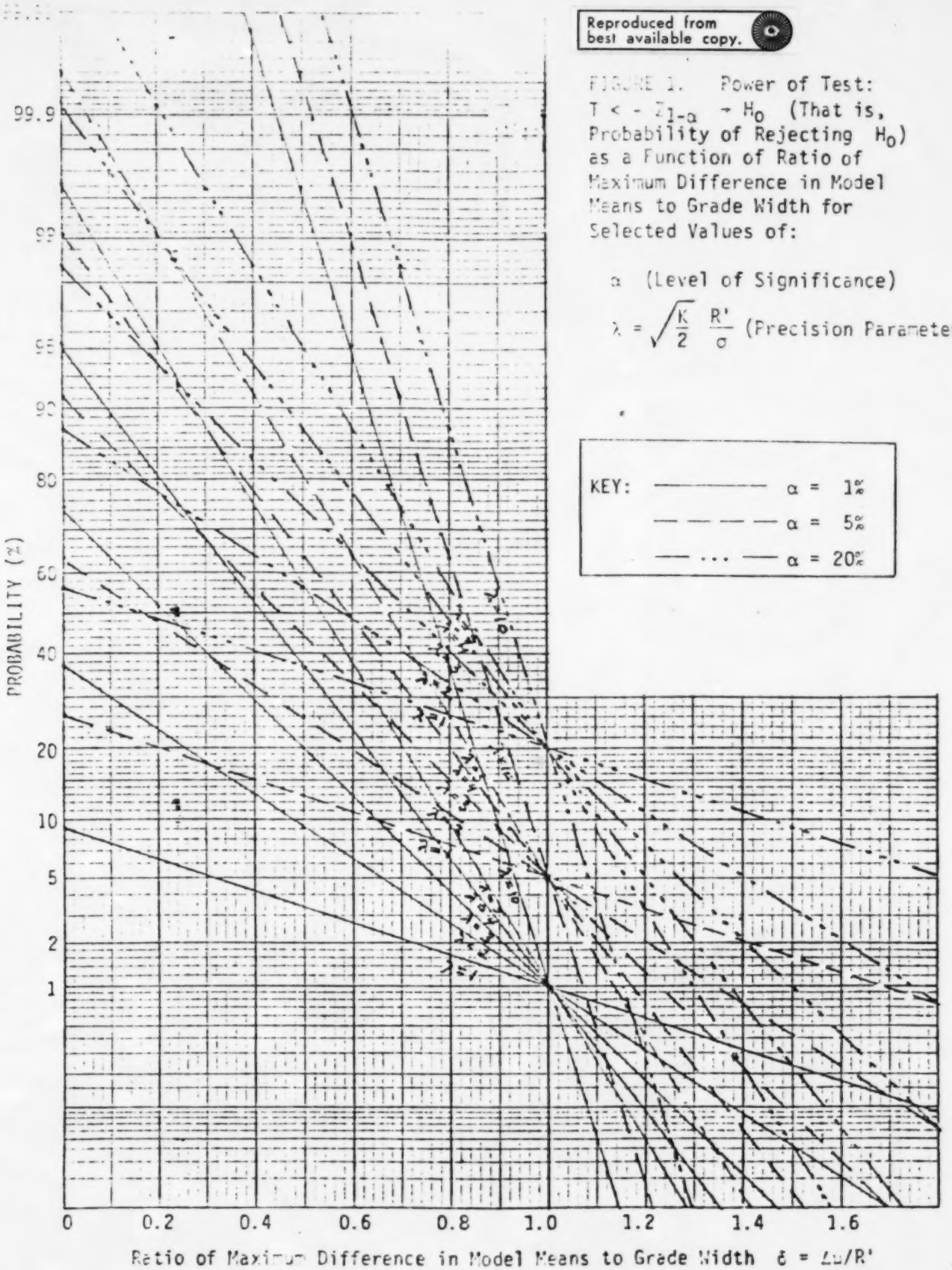

FIGURE 1. Power of Test: $T < -Z_{1-\alpha} \rightarrow H_0$ (That is, Probability of Rejecting H_0) as a Function of Ratio of Maximum Difference in Model Means to Grade Width for Selected Values of:

α (Level of Significance)

$\lambda = \sqrt{\frac{K}{2}} \frac{R'}{\sigma}$ (Precision Parameter)

Fixing, for the moment, on the case $\alpha = 10\%$ (which is perhaps as high as one might want to go as an acceptable level of significance), we ask the question: what level of precision is required to achieve at least 90% probability of acceptance of H_1 if $\Delta\mu$ is no greater than 75% of R'? Examination of Figure 1 yields $\lambda \sim 10$. This numerical illustration is not at all unreasonable as a set of practical performance requirements to impose on the oil grading procedure. Recapitulating: if $\Delta\mu \geq R'$ we want to reach the decision that with probability at least 90% that model variability is unacceptably large, and if $\Delta\mu \leq 0.75\ R'$ we want to reach the decision with probability at least 90% that model variability is within acceptable limits.

Now then, how severe is the requirement $\lambda = 10$? If we set: $R = 2.5\%$ (which would be the case if the current draft procedure made a slight alteration in Grade B-C boundary from 1.01 to 1.005) and $\rho = 1$, then $\sigma/\sqrt{K}$ must equal 0.18% to satisfy $\lambda = 10$. The following combinations of K and σ would meet this requirement:

K	σ
1	0.18%
4	0.35%
16	0.71%
64	1.41%
1024	5.66%

Suppose we arbitrarily relaxed the precision requirement to $\lambda = 5$. This implies at least 90% power for $\Delta\mu \leq 0.49\ R'$ which is of course appreciably poorer performance, but perhaps acceptable. (The impact of such a relaxation is that more real situations of truly acceptable model-to-model variability are likely to end up with the oil left ungraded.) Now, the requirement on $\sigma/\sqrt{K}$ is 0.35% and a comparable set of K, σ pairs is

K	σ
1	0.35%
4	0.71%
16	1.41%
256	5.66%

Taking the ASTM estimates of $\sigma = 1.6\%$ and 6.0% for the non-carryover and carryover procedures, we conclude that 320 cars and over 5000 cars, respectively, should be tested, under $\lambda = 10$ for each candidate oil. These numbers reduce to 80 and nearly 1300 if λ is relaxed to 5. Even with these large numbers there is no assurance, particularly so for the smaller λ, that a final grade would be achieved when model variability is small.

APPENDIX B

ACCOMMODATING SPURIOUS OBSERVATIONS IN FUEL ECONOMY MEASUREMENTS

Replicated fuel economy measurements at a particular mileage point introduce the potential of reducing random errors through averaging and also providing some degree of protection against errors caused by spurious observations. Since measurement costs are considerable, a limit is imposed on the maximum permitted number of replications. For purposes of the present discussion, we set this limit at 3. If higher limits are of interest, then the analysis presented below would have to be appropriately modified.

The Measurement Model

We denote the (unknown) true fuel economy of a vehicle at a given mileage point by μ, and three successive fuel economy measurements at that mileage by X_1, X_2, X_3.* When none of these measurements are spurious, the X_i are assumed statistically independent and each normally distributed about μ with variance σ^2. It is assumed that either the standard deviation σ itself is known or $\sigma = \mu\tau$ where τ (the coefficient of

---

* Mileage accumulation caused by the experimental procedure itself is neglected and the vehicle is assumed to be in the same state at initiation of each measurement.

variation) is known. These alternative models are designated I and II respectively for future discussion. The assumption of known measure of dispersion makes sense in the context of a standard experimental procedure for which a large body of prior experience is available.

If one could be certain that all observations are "good," i.e., none are spurious, then a best estimator of μ is the sample mean $\bar{X} = (X_1 + \ldots + X_n)/n$, and the choice of one, two, or three (or more) measurements is a relatively straightforward trade-off between cost and estimation error. Specifically, the mean square error (MSE) of this estimate (which is equal to its variance because of unbiasedness) is inversely proportioned to n while cost increases approximately linearly with n. The actual choice of n would be determined by absolute constraints on acceptable costs and acceptable errors and/or conceivably by minimization of total "cost" which incorporates a dollar-equivalent error component.

Modelling the Spurious Observation

The real problem, however, is compounded by the possible presence of spurious observations. By the very nature of the concept "spurious," we

should not presume to be able to give it a probabilistic characterization. However, in conformity with common practice, we assume that a spurious observation is also normally distributed with variance σ^2, but with mean shifted away from μ by an arbitrary number of standard deviation units to $\mu + b\sigma$. (A deterministic reading of $X_i = \mu + b\sigma$, i.e., all probability mass concentrated at $\mu + b\sigma$, but with μ and $b\sigma$ separately unknown, would probably best capture the essence of the "spurious" observation; the introduction of the random normal spread about $\mu + b\sigma$ has only a secondary averaging effect over some limited range of severity of spuriousness.)

An additional aspect of spurious observations is their frequency of occurrence. Under good experimental controls, which we assume applies to the problem at hand, spurious observations should be relatively rare events. More specifically, what is really desired is that the likelihood of <u>more than one</u> spurious reading among three measurements be negligibly small. Such a property appears as a natural consequence of independence in the occurrence of relatively rare spurious readings. For example, if each measurement has an independent probability, $\lambda < 0.05$, of giving rise to a spurious reading then the probabilities of zero (P_0), one (P_1), and more than one ($P_{>1}$) spurious observations among three measurements are bounded by:

$$P_0 > 1 - 3\lambda > 0.85$$

$$P_1 = 3\lambda(1 - \lambda)^2 < 0.14$$

$$P_{>1} = 3\lambda^2(1 - \frac{2}{3}\lambda) < 0.0075$$

Thus, on the average, fewer than one out of every 130 measurement triples would harbor multiple spurious observations. Our concern with limiting multiple spurious occurrences stems from the fact that among three observations they would be essentially uncorrectable. We therefore assume for purposes of our analyses that multiple spurious observations will not occur.

General Strategies for Dealing with Spurious Observations

A group of measurements is to be used to estimate a parameter. How would a spurious observation be distinguishable in the measurement set? Basically, it is likely to appear as an "outlier," that is, an unexpectedly extreme observation or as an unexpectedly large discrepancy between two measurements.

A traditional approach for dealing with spurious observations is to explicitly test for and identify outliers by a prescribed procedure, and then to eliminate them as "bad" data points. The problem with this approach is that good (non-spurious) observations occassionally manifest as outliers while, conversely, a small to moderate spurious shift, $b\sigma$, may frequently not be distinguishable.

A more sophisticated approach is to devise estimaters, based on all the data, with the property of being relatively insensitive to contamination from spurious observations, yet without explicitly having to decide which if any observation is spurious. Since the ultimate objective is accurate parameters estimation (in our case the mean of a normal distribution) there is no particular need to reach a yes/no decision as to whether a given observation is good or bad. Rather, all observations can be utilized, with the weight each carries dependent on some measure of its concordance with the total data set. Such an approach is generally termed "accommodation" to spurious effects or robust estimation.

Specific Mean Estimators

A classic paper treating mean estimation from n normally distributed observations in which at most one may have a spurious shift is by Anscombe.[1] The estimator proposed for the case of known σ is:

$$\hat{\mu}_A = \begin{cases} \bar{X} & |Zm| < c\sigma \\ \bar{X} - \frac{Zm}{n-1} = \frac{1}{n-1} \sum_{i \neq m} X_i & |Zm| \geq c\sigma \end{cases}$$

where $\bar{X} = \frac{1}{n} \Sigma X_i$ (sample mean)

$Z_m = X_m - \bar{X}$ is the m^{th} deviation from the sample mean with the property of being largest in absolute value

$c > 0$ (criterion to be selected).

Thus, the sample mean and all deviations are computed. If the maximum absolute deviation Z_m is sufficiently small the sample mean is accepted as the estimate; if Z_m exceeds a specified criterion, then X_m is excluded and the reduced sample mean is used. This approach is tantamount to identifying and rejecting X_m as an outlier when it is too extreme. An analogous estimator applicable to the case in which the coefficient of variation τ is known (Model II) could be formulated as:

$$\hat{\mu}_{A,II} = \begin{cases} \bar{X} & |Zm| / \bar{X} < c\tau \\ \bar{X} - \frac{Zm}{n-1} = \frac{1}{n-1} \sum_{i \neq m} X_i & |Zm| / \bar{X} > c\tau \end{cases}$$

and makes sense if τ is reasonably small, say $< 10\%$.

Veale and Huntsberger[2] developed a modification of Anscombe's estimator which they have shown to be at least as good and in some situations have superior performance. It is given by:

$$\hat{\mu}_{VH} = \begin{cases} \bar{X} & |Zm| < c \\ \bar{X} - \frac{Zm}{n-1} \left[\frac{Zm^2}{\frac{n-1}{n}\sigma^2 + Zm^2} \right] & |Zm| \geq c \end{cases}$$

Note that for $|Z_m| < c\sigma$, $\hat{\mu}_{VH}$ is identical to $\hat{\mu}_A$, and that for very large $|Z_m|$ it is approximately equal to $\hat{\mu}_A$, i.e., the reduced sample mean. However, for intermediate values of $|Z_m|$ an intermediate weight is given to X_m (the observation with largest absolute deviation), the

extent of which increases with decreasing $|Z_m|$ until the condition $|Z_m| < c\sigma$ is reached, at which point an abrupt transition occurs and "full" weight is given to X_m as an element of the sample mean. As with Anscombe, an analogous Veale-Huntsberger estimator for Model II is constructed by replacing σ by $\bar{X}\tau$, i.e.,

$$\hat{\mu}_{HV,II} = \begin{cases} \bar{X} & \frac{|Zm|}{\bar{X}} < c\tau \\ \bar{X} - \frac{Zm}{n-1}\left[\frac{Zm}{\frac{n-1}{n}\bar{X}^2\tau^2 + Zm^2}\right] & \frac{|Zm|}{\bar{X}} \geq c\tau \end{cases}$$

Numerous other robust estimators have been proposed based on, e.g., maximum likelihood principle, order statistics, and rank order statistics.[3] Generally the focus has been on achieving good asymptotic performance for very large n. Since our concern is with $n \leq 3$ and there is no evidence of real advantage of these approaches in our context, we shall not consider them.

The estimators so far presented are clearly only meaningful for $n \geq 3$, and, with our restriction, only for $n = 3$. What if there is a strong incentive because of high measurement cost to try to reduce the number of observations to 2*, at least to achieve such a reduction in a significant proportion of cases? Are any effective approaches available?

---

* There is no hope of dealing objectively with spurious observations if only a single measurement is taken; hence the alternative of n=1 is not considered.

Desu, Gehan, and Severo[4] have considered such a problem and propose a two-stage estimation procedure wherein only two measurements are taken at the first stage. If these are sufficiently close to each other, their mean is accepted as the estimate. Otherwise, a third measurement is taken. Their procedure in the latter case is not relevant to our problem since they then assume the third observation to be non-spurious. However, we can incorporate the concept of a two-stage procedure into the Veale-Huntsberger estimator as follows, for n=3:

$$\hat{\mu}_K = \begin{cases} \dfrac{X_1 + X_2}{2} & |X_1 - X_2| < d\sigma \\ \bar{X} & |X_1 - X_2| \geq d\sigma \;;\; |Zm| < c\sigma \\ \bar{X} - \dfrac{Zm}{2}\left[\dfrac{Zm^2}{\frac{2}{3}\sigma^2 + Zm^2}\right] & |X_1 - X_2| \geq \sigma \;;\; |Zm| \geq c\sigma \end{cases}$$

Again, as before, for Model II:

$$\hat{\mu}_{K,II} = \begin{cases} \dfrac{X_1 - X_2}{2} & \dfrac{|X_1 - X_2|}{(X_1 + X_2)/2} < d\tau \\ \bar{X} & \dfrac{|X_1 - X_2|}{(X_1 + X_2)/2} \geq d\tau \;;\; \dfrac{|Zm|}{\bar{X}} < c\tau \\ \bar{X} - \dfrac{Zm}{2}\left[\dfrac{Zm^2}{\frac{2}{3}\bar{X}^2\tau^2 + Zm^2}\right] & \dfrac{|X_1 - X_2|}{(X_1 + X_2)/2} \geq d\tau \;;\; \dfrac{|Zm|}{\bar{X}} \leq c\tau \end{cases}$$

The promise of this two-stage estimator is that it may be able to retain a good deal of the robustness of performance of the Veal-Huntsbeyer estimator under possible spurious contamination while at the same time achieving an appreciable reduction in the average number of measurements conducted per estimation.

For comparison purposes, we consider two additional estimators, the sample mean of three observations, $\hat{\mu}_S$, and the staged procedure $\hat{\mu}_E$ described in a draft EPA Recommended Practice relating to engine oil fuel efficiency grading[5], except that it is modified to be terminated at three rather than four measurements. These are defined as follows:

$$\hat{\mu}_S = \hat{\mu}_{S,II} = \bar{X}$$

$$\hat{\mu}_E = \begin{cases} \dfrac{X_1 + X_2}{2} & |X_1 - X_2| < d\sigma \\ \dfrac{1}{2}\displaystyle\sum_{\substack{i=1 \\ i\neq m}}^{3} X_i & |X_1 - X_2| \geq d\sigma \end{cases}$$

$$\hat{\mu}_{E,II} = \begin{cases} \dfrac{X_1 + X_2}{2} & \dfrac{|X_1 - X_2|}{(X_1 + X_2)/2} < d\tau \\ \dfrac{1}{2}\displaystyle\sum_{\substack{i=1 \\ i\neq m}}^{3} X_i & \dfrac{|X_1 - X_2|}{(X_1 + X_2)/2} \geq d\tau \end{cases}$$

where X_m is the measurement with largest absolute deviation from the sample mean. Note in the case of $\hat{\mu}_E$ and $\hat{\mu}_{E,II}$, that when a third observation is taken, the estimate corresponds to the "best two out of three" rule.

We proceed next to consider, in detail, the performance of the estimators that have been defined.

Estimator Performance

An individual estimate of $\hat{\mu}$ of μ (the true fuel efficiency of the test car) will be in error by an amount $\hat{\mu}-\mu$. A useful, as well as mathematically convenient, statistical measure of esuch errors is the expected value of the square of the error, also called the mean square error, MSE:

$$MSE\ (\hat{\mu}) = E\ [\ (\hat{\mu}-\mu)^2\ ].$$

The practical utility of MSE stems from the fact that the weight given to large errors is enhanced by squaring, corresponding to the subjective notion that the seriousness or disutility ascribable to an error grows with increasing rapidity as its magnitude increases.

In general, one can relate MSE to the bias and variance of an estimator

$$\text{MSE}(\hat{\mu}) = \left(E[\hat{\mu}] - \mu\right)^2 + \text{Var}[\hat{\mu}]$$

The first term is the square of the bias. If an estimator is unbiased, i.e., if its expected value is equal to the parameter μ being estimated, then $\text{MSE}(\hat{\mu}) = \text{Var}[\hat{\mu}]$. However, in the face of contamination by spurious observations, estimators will generally be biased (in a manner depending on the direction and magnitude of the mean shift), so it is inappropriate to use variance alone as the measure of performance.

In comparing MSE of $\bar{X}$, $\hat{\mu}_A$, and $\hat{\mu}_H$, we normalize by MSE $(\bar{X})$ under condition of no spurious observation, equal to $\sigma^2/3$ for $n=3$. This represents optimum performance in that $\bar{X}$ is then an unbiased and efficient estimator. Figure 1 graphs $3\text{MSE}(\bar{X})/\sigma^2$, $3\text{MSE}(\hat{\mu}_A)/\sigma^2$ (for $c=2.460$) and $3\text{MSE}(\hat{\mu}_{VH})/\sigma^2$ (for $c=2.404$) as functions of b, the magnitude of spurious shift from the mean.* With the normalization

---

* The numerical values are taken from calculations by Veale and Huntsberger.[2] However, it should be remarked that the calculations were based on the unjustified assumption that the spurious contaminant may be identified with the measurement having the largest absolute deviation from the sample mean. This leads to erroneously low values for MSE at intermediate b-values. For example, whereas they compute $3 \cdot \text{MSE}/\sigma^2 = 2.94$ for $b=4$, $c=2$, a direct Monte Carlo simulation of this case yielded 3.41 (with simulation standard deviation ~ 0.05). Nevertheless, the results plotted should be reasonably valid for relative comparison of estimators and for discussion of qualitative features.

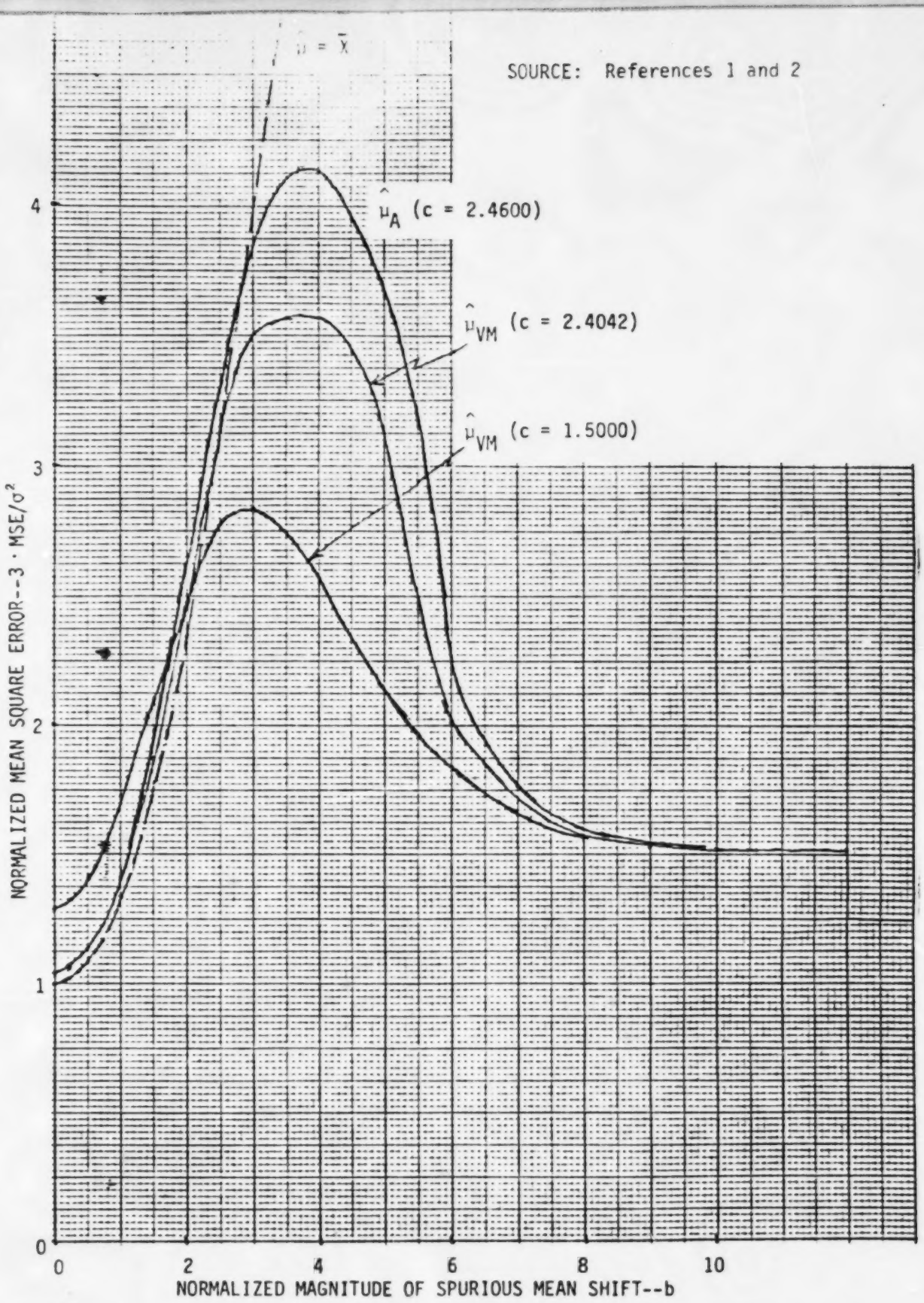

FIGURE 1. Comparison of Estimators of Mean of Normal Distribution Using Three Observations, When One Observation has Spurious Mean Shift

defined above, the ordinate value of 1 is the ideal limit, achievable only if no spurious observation is present (b=0). Note that while $MSE(\bar{X})$ is, by definition, equal to 1 for b=0, it increases rapidly and without bound with increasing b, demonstrating lack of robustness. $MSE(\hat{\mu}_A)$ and $MSE(\hat{\mu}_{VH})$, on the other hand, while paying a modest premium of 4% relative to ideal at b=0, increase to maximum values of 3.5 and 4.25, respectively, times the ideal at b in the vicinity of 3-4 and then diminish again, both approaching a limiting value of 1.5 as $b \to \infty$. The latter limit corresponds to the reduced sample mean using the two "good" observations and its associated variance of $\sigma^2/2$, compared to $\sigma^2/3$ for $\bar{X}$. The particular c-values for $\hat{\mu}_A$ and $\hat{\mu}_{VH}$ were selected to yield a 4% premium at b=0. As evident from the additional plot of $\hat{\mu}_{VH}$ for c=1.5, lower c-values can achieve better protection against intermediate values of b, but at the expense of an increased premium at b=0. Conversely, larger c-values would lower the premium but at the expense of reduced protection for intermediate values of b. The curve maxima would increase and shift to higher b. Note finally the modestly superior performance of $\hat{\mu}_{VH}$ relative to $\hat{\mu}_A$.

In order to determine the performance of the $\hat{\mu}_K$ and $\hat{\mu}_E$ estimators, Monte Carlo computer simulations were performed for selected criteria values of d and c and for selected spurious shift magnitudes b occurring in the 1st, 2nd, or (potentially) 3rd measurement with equal

likelihood. Results are depicted in Figures 2, 3, and 4 with straight line segments connecting the different b-values for each specified estimator. Clearly the fine scale dependence on b is only roughly approximated by linear interpolation, but the same qualitative behavior of MSE peaking at some intermediate b-value, as exhibited in Figure 1, occurs also with these estimators.

Also shown is the mean number of measurements per estimate, N, which depends on d and b and ranges between 2 and 3. N clearly decreases with increasing d. On the other hand, for fixed d, N increases with increasing b. If we denote the values of N at $b=0$ and $b \to \infty$ by N_0 and N_∞, respectively, then it can be shown that $N_\infty = 2 + N_0/3$.

Because of the reduced numbers of observations, the normalizing factor applied to MSE is σ^2/N rather than $\sigma^2/3$. This corresponds approximately to the variance of a randomized mix of the 2-sample and 3-sample means as follows:

$$\hat{\mu}_R = \left\{ \begin{array}{ll} (X_1+X_2)/2 & \text{with probability } 3-N \\ (X_1+X_2+X_3)/3 & \text{with probability } N-2 \end{array} \right\} \text{ independent of } X_1, X_2, X_3$$

When no contamination is present, and under the constraint that the mean number of observations per estimate = N $(2 \leq N \leq 3)$, $\hat{\mu}_R$ is an unbiased

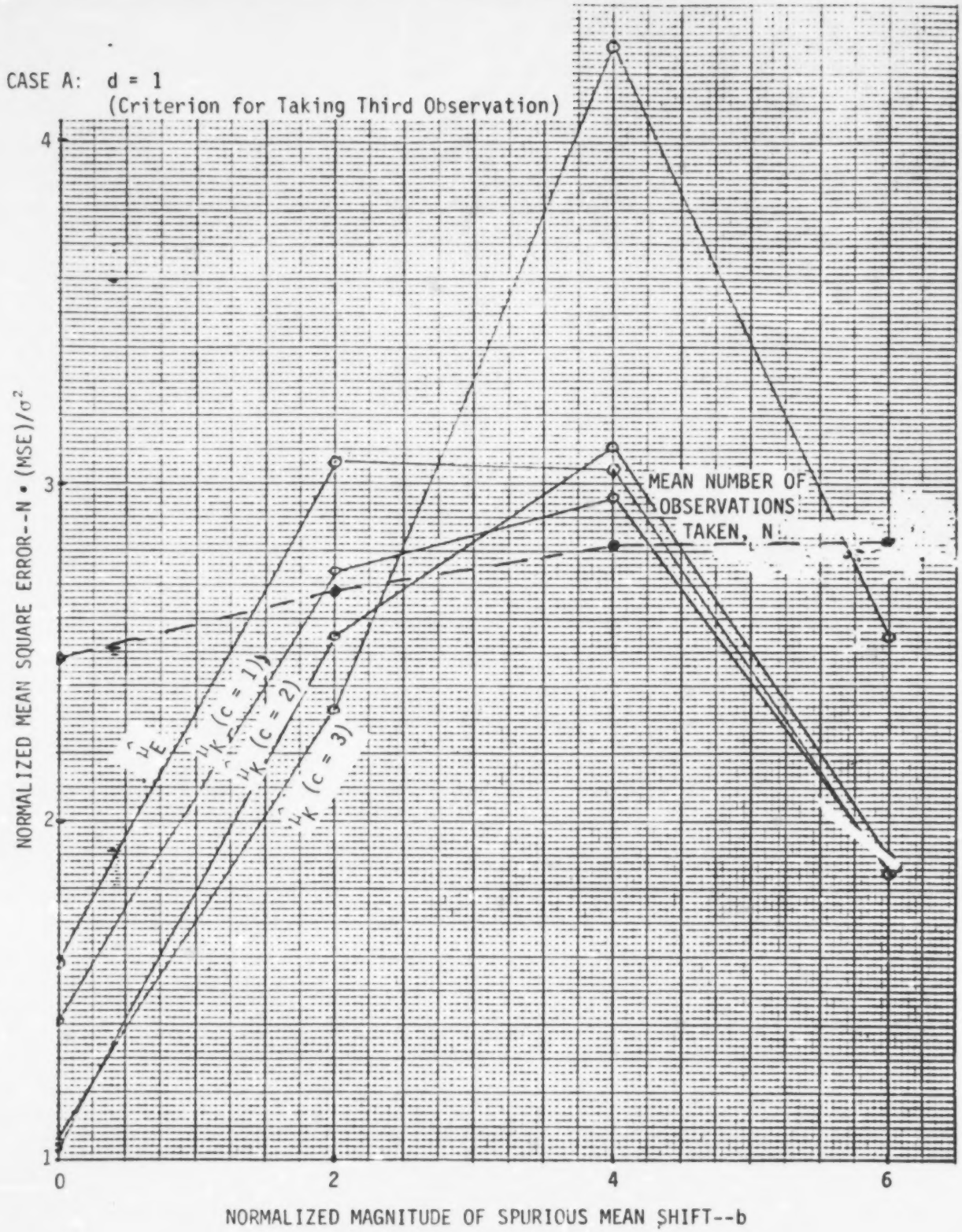

FIGURE 2. Performance of 2-Stage Estimators of Mean of Normal Distribution Using Two or Three Observations, When One Observation has Spurious Mean Shift

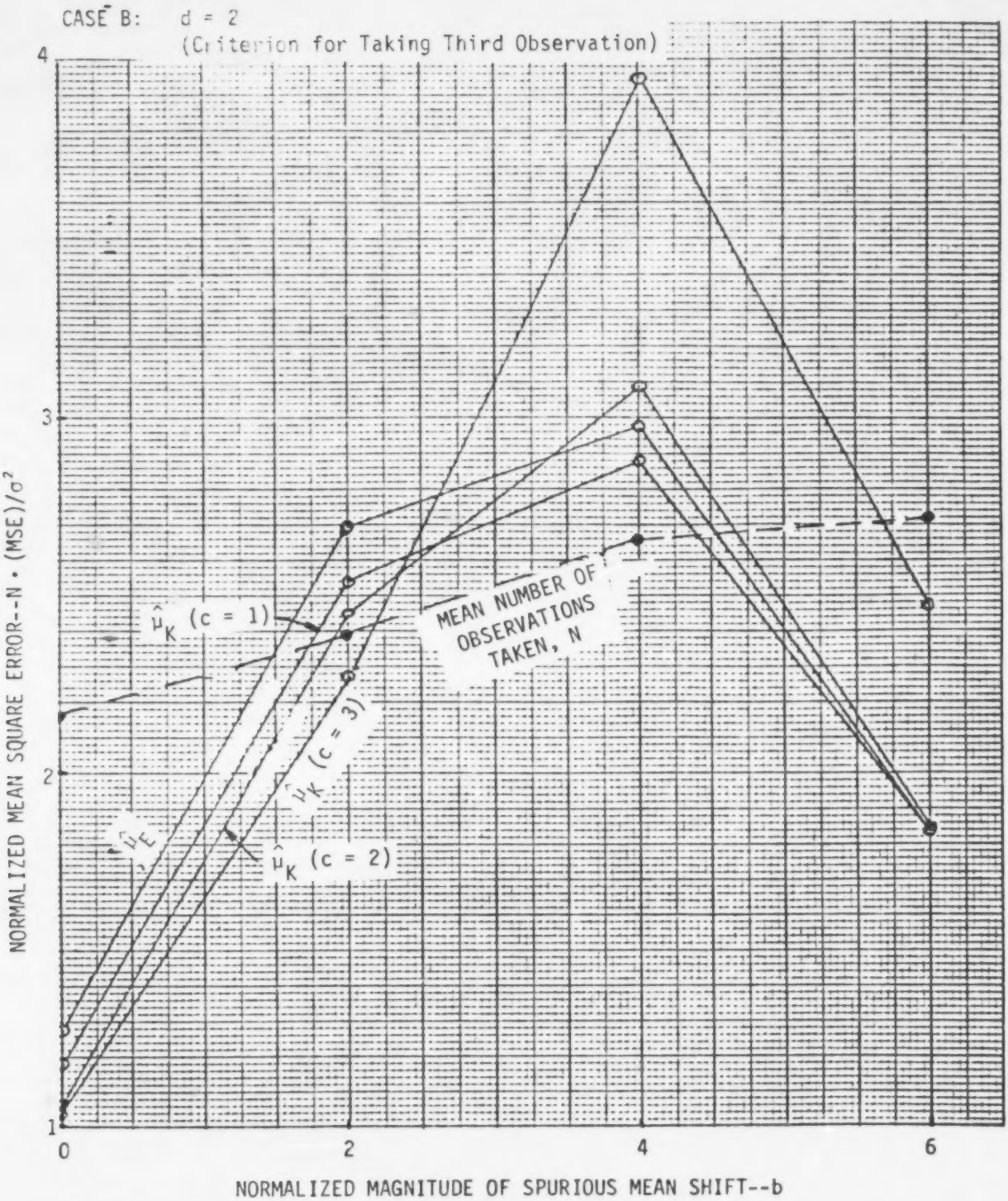

FIGURE 3. Performance of 2-Stage Estimators of Mean of Normal Distribution Using two or Three Observations, When One Observation has Spurious Mean Shift

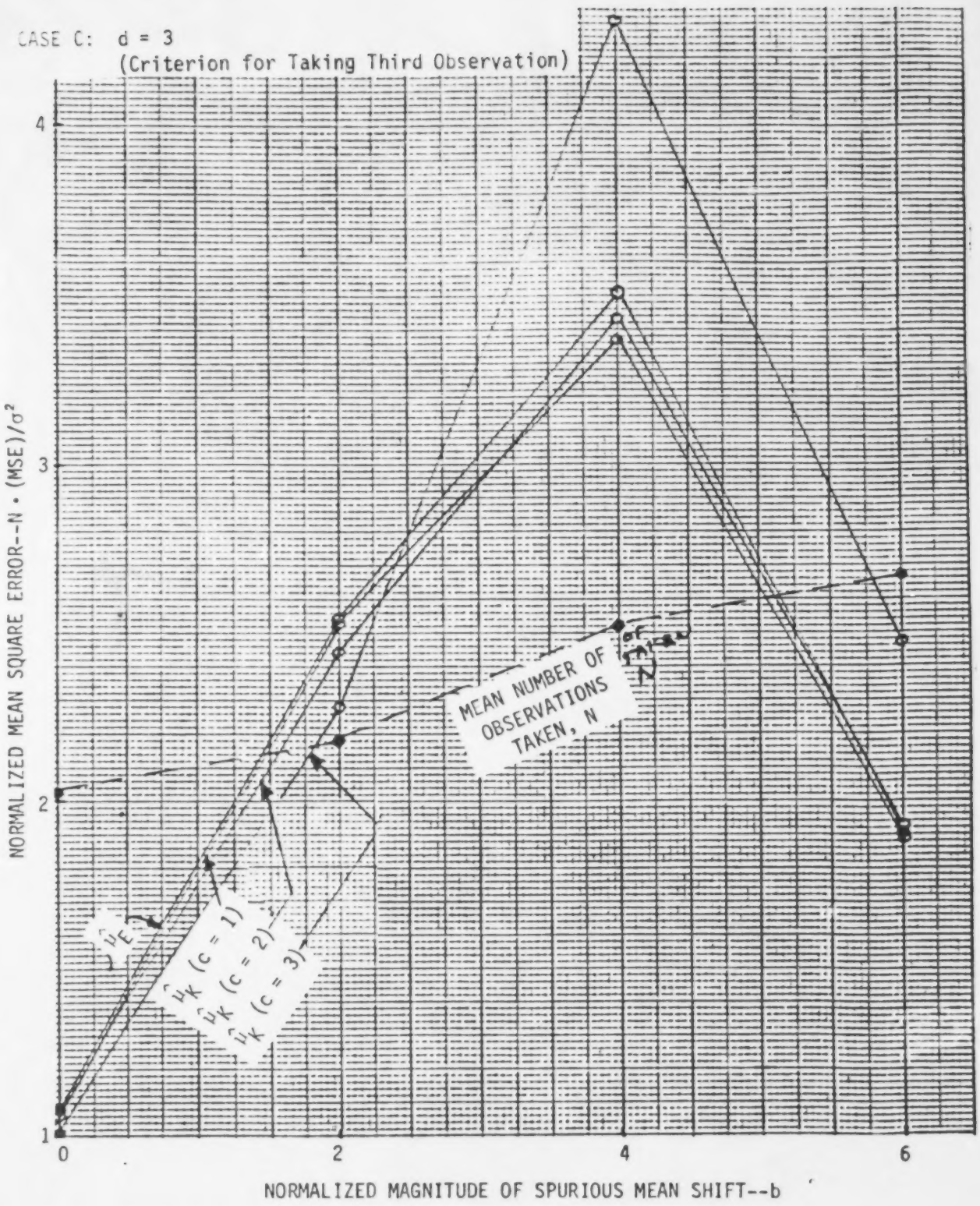

FIGURE 4. Performance of 2-Stage Estimators of Mean of Normal Distribution Using Two or Three Observations, When One Observation Has Spurious Mean Shift

efficient estimator with variance approximately equal to σ^2/N.*

Examination of Figures 2, 3, and 4 suggests that it should be possible to determine values of c and d such that $MSE(\hat{\mu}_K)$ will be below $MSE(\hat{\mu}_E)$ for all possible magnitudes of spurious mean shift b. In particular, the condition b=0, which corresponds to no contamination, is crucial since that is expected to be the case most of the time. Hence, great emphasis should be given to minimizing MSE at that point. Note that $\hat{\mu}_E$ performance in that region begins to become acceptable only for $d \geq 3$, but then the associated peak MSE at irtermediate b-values becomes relatively large.

Optimization of $\hat{\mu}_K$ with respect to choice of c, d requires specification of further information and definition of optimality since MSE depends on b. One possibility is to estimate a prior distribution for b and then to minimize the expected value of MSE (Bayesian approach). An alternate approach, which depends only on knowing the probability λ that any given measurement is contaminated,

---

* Strictly speaking, the variance of $\hat{\mu}_R$ is $(5-N)\sigma^2/6$; hence that is the proper normalizing factor. However, the maximum discrepancy between $(5-N)/6$ and $1/N$ within the range $[2,3]$ is 4% (when $N = \sqrt{6} = 2.45$), so $1/N$ is a quite acceptable approximation.

would minimize

$$(1 - N_0\lambda) \cdot MSE(\hat{\mu}_K)_{b=0} + N_0\lambda \cdot \max_{b>0} \left\{ MSE(\hat{\mu}_K)_b \right\}.$$

This is a hybrid of minimax and Bayesian approaches. The first term is the contribution to MSE when there is no contamination. The second term is the maximum possible contribution to MSE when a spurious observation of unknown magnitude has been included. Although additional Monte Carlo simulation would be required to perform such an optimization with precision, the results on hand suggest that with λ of the order of .02 to .05 the solution would not be too far from c=d=2.

Notwithstanding the defect in the numerical results for $\hat{\mu}_{VH}$, we can attempt a rough bounding comparison of $\hat{\mu}_{VH}$ and $\hat{\mu}_K$. Specifically, compare these two with former specified by c=2.404 and the latter by d=1, c=2. Note that MSE at b=0 are comparable but that $\hat{\mu}_K$ has a lower peak MSE. While the true peak MSE of $\hat{\mu}_K$ may be higher at b-values slightly different from 4, it is likely to remain below the MSE peak of $\hat{\mu}_{VH}$ indicated in Figure 1. Furthermore the latter is probably significantly below true value because of the previously noted incorrect assumption.

REFERENCES

[1] F. J. Anscombe, "Rejection of Outliers," Technometrics, 2:123, 1960.

[2] J. R. Veale and D. V. Huntsberger, "Estimation of a Mean When One Observation May be Spurious," Technometrics, 11:331, 1969.

[3] V. Barnett and T. Lewis, Outliers in Statistical Data, John Wiley & Sons, New York, Pages 144-145, 1978.

[4] M. M. Desu, E. A. Gehan, and N. C. Severo, "A Two-Stage Estimation Procedure When There May be Spurious Observations," Biometrika, 61:593, 1974.

[5] Environmental Protection Agency, "EPA Recommended Practice for Evaluating, Grading, and Labeling the Fuel Efficiency of Motor Vehicle Engine Oils," No Number, No Date.

APPENDIX C

"BEST TWO OUT OF THREE" PROCEDURES

An issue related to the repeatability test is "the fallacy of the best two out of three," which was documented in a series of papers by Youden[1,2] and Lieblein.[3] "The best two out of three" refers to a common practice in the chemical laboratory of taking a third determination "to indicate which of the other two is more likely to be off the mark." If two of the three measurements are in close agreement, the experimenter discards the remaining one as representing some gross error which renders it invalid. The two authors showed: (1) that the spacing between the most discrepant observation and its nearest neighbor is often many times the spacing between the two closest observations, even where all represent valid estimates of the same (population) value, (2) that use of the closest pair of observations causes the experimental error (the underlying standard deviation) to be underestimated, and (3) that the mean of the closest pair out of three is subject to larger variability than the mean of a fixed sample of two determinations. (It is clear that its variance is larger than that of the mean of a fixed sample of three observations.)

It is not difficult to illustrate the "fallacy." Monte Carlo experiments with three normal populations, all with the same mean, showed[4] that the standard error of the mean of two readings chosen according to "the best two out of three" is approximately 0.8σ, where σ is the error variance common to all populations. This is to be compared with $\sigma/\sqrt{2} = 0.71\sigma$ for two observations chosen completely at random, and with $\sigma/\sqrt{3} = 0.58\sigma$ for three randomly chosen observations. The reason for this is the asymmetry frequently shown among the three observations.

---

[1] W. J. Youden, "The Fallacy of the Best Two Out of Three," NBS Technical Bulletin, 33: 77-78 (1949).

[2] W. J. Youden, "Sets of Three Measurements," The Scientific Monthly, 77: 143-147 (1953).

[3] J. Lieblein, "Properties of Certain Statistics Involving the Closest Pair in a Sample of Three Observations," Journal of Research of the National Bureau of Standards, 48: 255-268 (1952).

[4] The exact value, based on a mathematical derivation given in Lieblein's paper, is 0.7986σ. Our results, based on 1000 trials, yielded the estimate 0.7945σ.

Suppose that the values are placed in order, so that X_1, X_2, X_3 implies $X_1 \leq X_2 \leq X_3$. Let D represent the larger of the two spacings $(X_2 - X_1, X_3 - X_2)$ and let d represent the smaller. Then the ratio $Q = D/d$ is 1 when (X_1, X_2, X_3) are symmetrically arranged, and is large when two of the observations are close together compared with the distance to the third. Whenever Q is large, there is a tendency to compromise the averaging effect associated with random sampling, where large negative errors "cancel out" positive errors of the same magnitude. Instead, there is more of a tendency to base the estimate on pairs of observations "off in one corner." It is this tendency that inflates the standard error of the sample mean. But note that large Q's not only are associated with increased sampling errors; a sample with a large Q is most likely to convince an unsuspecting experimenter that one of the observations is an outlier.

The probability of obtaining a large Q is surprisingly high, even where there are no outliers. In our Monte Carlo experiments, it was shown that $P(Q \geq 5) \sim .30$. Indeed, the same experiments showed that $P(Q \geq 10) \sim .15$. In other words, it is not uncommon to see two readings close together, and a third some distance away (in fact, five or ten times farther away), due to chance alone.

Knowing this fact, it might be expected that Q would provide a poor discriminator of outliers. Monte Carlo experimentation showed that this was indeed the case. Suppose that one of the three populations (say population 1) is shifted away from the others. Explicitly, let μ_i and σ_i be the mean and standard deviation of the i^{th} population $(i = 1, 2, 3)$ and suppose

$$(1) \quad \sigma_i = \sigma \qquad (i = 1, 2, 3)$$

$$(2) \quad \mu_2 = \mu_3 = \mu$$

$$(3) \quad \mu_1 = \mu + \delta\sigma$$

Since the justification of "the best two out of three" is its ability to screen out outliers, it is of interest to know how sensitive the distribution of Q is to the size of δ and how often the procedure would succeed in eliminating the observation from population 1 from the pair which goes into the sample estimate. Table C-1 shows the results of our experiments for various sizes of δ.

Let $\overline{U}$ be the mean of the two observations that are closest together, i.e., the pair that determines the denominator of Q. The value Q_1 of Table C-1 is by definition

Table C-1

BEHAVIOR[5] OF "BEST TWO OUT OF THREE" SAMPLES
FOR SELECTED VALUES OF δ

δ	$E(\bar{U})$	$SD(\bar{U})$	$P(5 \leq Q < 10)$	$P(Q \geq 10)$	Q_1
0	μ	0.80σ	.147	.149	0.67
0.5	$\mu + .14\sigma$	0.82σ	.149	.157	0.64
1.0	$\mu + .29\sigma$	0.86σ	.138	.158	0.59
2.0	$\mu + .40\sigma$	0.98σ	.157	.150	0.39
2.5	$\mu + .37\sigma$	1.02σ	.156	.161	0.30
3.0	$\mu + .31\sigma$	1.03σ	.177	.175	0.21
5.0	$\mu + .06\sigma$	0.86σ	.212	.272	0.03

[5] Based on Monte Carlo experiments; N = 1000 trials.

Q_1 = P ($\bar{U}$ Contains the Population 1 Observation)

Note that $Q_1 = 2/3$ when $\delta = 0$. Table C-1 also shows the behavior of $E(\bar{U})$ and $SD(\bar{U})$ with δ.

The table shows that there is a substantial probability of including the outlier unless its distribution is centered 5σ or more away from the other two population means. This is associated with the behavior of the distribution of Q; the latter is insensitive to δ until δ is at least 5. Because of the positive probability of including the offset population, $\bar{U}$ becomes a biased estimator. The amount of bias is determined by the probability of including the observation from population 1, and the distance δ between population 1 and the others. Up to about $\delta = 2$, the size of the distance dominates, and the amount of bias increases even though Q_1 is getting smaller. For larger δ, the decreasing size of Q_1 takes over, and the amount of bias decreases. A similar effect is observed with respect to $SD(\bar{U})$, although the decrease doesn't begin until $\delta \sim 3$. As δ increases and Q_1 approaches zero, $SD(\bar{U})$ will decrease and will asymptotically approach $\sigma/\sqrt{2}$.

The application to fuel economy testing is straightforward. If a spurious observation is included among three, there is little chance of weeding it out by taking the closest pair unless the outlier is about 5σ away. Thus judgment on the efficacy of the procedure depends to some extent on the size of σ. Generally, we have been concerned with the range $0.75\% \leq \sigma \leq 2\%$, i.e., .15 mpg $\leq \sigma \leq$ .4 mpg for a 20 mpg model. This indicates that "the best two out of three" could be effective if an outlier is over 0.75 mpg away from the others, and that fine a discrimination would be possible only if σ = .15. If the σ = .4 mpg estimate were realistic, then there would be little chance of eliminating the outlier from a triple of observations unless it were 2 mpg removed from the others.

Of course, the EPA proposed procedure is not "the best two out of three." It differs in that one is not committed <u>beforehand</u> to taking the third observation, but only does so when the difference between the first two exceeds a preassigned criterion. Perhaps more important, the "best two out of three" <u>forces</u> a discard of one of the observations, rather than allowing all three to be used if (as in the repeatability test) they are mutually close. Nevertheless, analysis of the "best two out of three" sheds light on the frequency which widely divergent, <u>non</u>-spurious observations can appear in the same small sample. Since such occurrences will tend to cause discards of good observations even when a preassigned criterion is used, these conclusions are pertinent to analysis of the repeatability test.

APPENDIX D

BEHAVIOR OF THE EPA REPEATIBILITY TEST

The repeatability test of the EPA proposed procedure was investigated, by means of Monte Carlo experimentation, to estimate the variability of fuel economy estimates yielded by the procedure, and to estimate the probabilities of requiring the second and third stages.

A preliminary approximation was available for the probability of advancing to the second stage. Note that

$$A_{12} = \frac{|X_1 - X_2|}{\frac{1}{2}(X_1 + X_2)} = R/\bar{X},$$

where R is the sample range. Taking $\bar{X} = 20$ mpg, for example,

$$P(A_{12} > .021) \sim P(R > .42).$$

Table D, p. 614 of Duncan[1] yields $E(R) = 1.128\sigma$, $SD(R) = 0.853\sigma$. When $\sigma = 0.15$ mpg (=.0075(20)), the procedure is designed so that $P(R > .42) = .05$. Suppose σ has a different value, say 1% of the mean ($\sigma = 0.2$ mpg). Then $\mu_R \sim .226$, $\sigma_R \sim .085$, and $.42 \sim \mu_R + 2.3\,\sigma_R$. Then[2] $P(R > .42)$ could not be large. On the other hand, if $\sigma = 2\%$ of the mean, i.e., $\sigma = .4$ mpg, then $.42 \sim \mu_R$ and thus $P(R > .42)$ would be near 0.5 for that σ.

A summary of the results of the Monte Carlo experiments is given in Table D-1. No spurious observations were present in these runs. It can be seen that the rough conjectures above were borne out. With $\sigma = 1\%$ of the mean, it was found that the estimate was based on the first two observations in 88% of the trials when there were no spurious observations; i.e., the third observation was employed 12% of the time. Under the same conditions, but with $\sigma = 2\%$ of the mean, the third observation was required in 48% of the trials. Thus only in those cases where σ is larger than anticipated, is there substantial probability of going beyond the first stage. When $(\sigma/\mu) = .75\%$, the design condition, the probability of going to the second stage is about 5%, as specified.

---

[1] A. J. Duncan, Quality Control and Industrial Statistics, R. D. Irwin, Inc., 663 pp, 1952.

[2] Chebychev's Inequality indicates $P(|R| > \mu_R + 2.3\,\sigma_R) < 0.19$.

Table D-1

MONTE CARLO EXPERIMENTS WITH THE REPEATABILITY TEST
(2.1% CRITERION)

σ (% OF μ)	NUMBER OF TRIALS	AVERAGE $(\hat{\mu}-\mu)/\mu$ (%)	$SD(\hat{\mu})/\sigma$	% TRIALS ENDING AFTER		
				2 OBS.	3 OBS.	4 OBS.
0.75	7500	+.001	.7625	95.28	4.72	0.
1.00	2500	+.027	.7667	86.32	13.60	0.08
1.25	2500	+.024	.7685	77.08	22.24	0.68
2.00	2500	-.021	.7806	52.36	39.80	7.84

When σ is in the range anticipated in establishing the 2.1% criterion, there are almost no cases in which the fourth observation is required. Indeed, even if employed in a "off-design" situation, such as $\sigma = .02\mu$, the fourth observation is required only 8% of the time. Thus the inclusion of the third stage in the procedure has very little effect on its results.

The estimates of μ obtained under the various sizes of σ were all substantially "on target." The deviations of those estimates from the true value are given in Table D-1, expressed as a percent of the true value. The differences shown represent random variability, since the estimator is unbiased when no spurious observations are present.

The procedure does share with the "best two out of three" the property of yielding estimators that are more variable than fixed samples of size two. The standard deviations obtained, all within the range $(.75\sigma, .79\sigma)$ are to be compared with $.7071\sigma$ for the fixed sample of size two. Note, in comparison, that the standard deviation among estimates yielded by the "best two out of three" is approximately 0.80σ when no spurious observations are present. Some tendency is apparent for the standard deviation to grow in disproportion to the growth in σ, but it is not clear whether or not this is a chance phenomenon.